

11-12-2015


Robust Modeling and Predictions of Greenhouse Gas Fluxes from Forest and Wetland Ecosystems

Khandker S. Ishtiaq

Florida International University, kisht001@fiu.edu

DOI: 10.25148/etd.FIDC000186

Follow this and additional works at: <https://digitalcommons.fiu.edu/etd>

 Part of the [Bioinformatics Commons](#), [Civil Engineering Commons](#), [Climate Commons](#), and the [Forest Sciences Commons](#)

Recommended Citation

Ishtiaq, Khandker S., "Robust Modeling and Predictions of Greenhouse Gas Fluxes from Forest and Wetland Ecosystems" (2015). *FIU Electronic Theses and Dissertations*. 2287.

<https://digitalcommons.fiu.edu/etd/2287>

This work is brought to you for free and open access by the University Graduate School at FIU Digital Commons. It has been accepted for inclusion in FIU Electronic Theses and Dissertations by an authorized administrator of FIU Digital Commons. For more information, please contact dcc@fiu.edu.

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

ROBUST MODELING AND PREDICTIONS OF GREENHOUSE GAS FLUXES
FROM FOREST AND WETLAND ECOSYSTEMS

A dissertation submitted in partial fulfillment of

the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

CIVIL ENGINEERING

by

Khandker Saqib Ishtiaq

2015

To: Interim Dean Ranu Jung
College of Engineering and Computing

This dissertation, written by Khandker Saqib Ishtiaq, and entitled Robust Modeling and Predictions of Greenhouse Gas Fluxes from Forest and Wetland Ecosystems, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

Leonard J. Scinto

Berrin Tansel

Hector R. Fuentes

Arindam Gan Chowdhury

Omar I. Abdul-Aziz, Major Professor

Date of Defense: November 12, 2015

The dissertation of Khandker Saqib Ishtiaq is approved.

Interim Dean Ranu Jung
College of Engineering and Computing

Dean Lakshmi N. Reddi
University Graduate School

Florida International University, 2015

© Copyright 2015 by Khandker Saqib Ishtiaq

All rights reserved

DEDICATION

I dedicate this dissertation to the memory of my mother, the best academician I have ever known.

ACKNOWLEDGMENTS

First and foremost I would like to express my sincere gratitude and thank to my major professor, Dr. Omar I. Abdul-Aziz for his continuous guidance, support, and mentorship for all these years at FIU. I would also like to thank the committee members, Dr. Leonard Scinto, Dr. Berrin Tansel, Dr. Arindam Chowdhury, and Dr. Hector Fuentes for their advice and insightful comments to this dissertation. This dissertation is funded by grants from national science foundation (NSF) and national oceanic and atmospheric administration (NOAA). The supports are gratefully acknowledged. I also sincerely acknowledge the support from FIU graduate school through a dissertation year fellowship.

Last but not the least, I would like to thank my family members especially my parents and my wife for their continuous moral support without which I could not have achieved this.

ABSTRACT OF THE DISSERTATION
ROBUST MODELING AND PREDICTIONS OF GREENHOUSE GAS FLUXES FROM
FOREST AND WETLAND ECOSYSTEMS

by

Khandker Saqib Ishtiaq

Florida International University, 2015

Miami, Florida

Professor Omar I. Abdul-Aziz, Major Professor

The land-atmospheric exchanges of carbon dioxide (CO₂) and methane (CH₄) are major drivers of global warming and climatic changes. The greenhouse gas (GHG) fluxes indicate the dynamics and potential storage of carbon in terrestrial and wetland ecosystems. Appropriate modeling and prediction tools can provide a quantitative understanding and valuable insights into the ecosystem carbon dynamics, while aiding the development of engineering and management strategies to limit emissions of GHGs and enhance carbon sequestration. This dissertation focuses on the development of data-analytics tools and engineering models by employing a range of empirical and semi-mechanistic approaches to robustly predict ecosystem GHG fluxes at variable scales.

Scaling-based empirical models were developed by using an extended stochastic harmonic analysis algorithm to achieve spatiotemporally robust predictions of the diurnal cycles of net ecosystem exchange (NEE). A single set of model parameters representing different days/sites successfully estimated the diurnal NEE cycles for various ecosystems. A systematic data-analytics framework was then developed to determine the mechanistic, relative linkages of various climatic and environmental drivers with the GHG fluxes. The analytics, involving big data for diverse ecosystems of the AmeriFLUX network, revealed robust latent patterns: a strong control of radiation-energy variables, a moderate control of temperature-hydrology variables, and a relatively weak control of aerodynamic variables on the terrestrial CO₂ fluxes.

The data-analytics framework was then employed to determine the relative controls of different climatic, biogeochemical and ecological drivers on CO₂ and CH₄ fluxes from coastal wetlands. The knowledge was leveraged to develop nonlinear, predictive models of GHG fluxes using a small set of environmental variables. The models were presented in an Excel spreadsheet as an ecological engineering tool to estimate and predict the net ecosystem carbon balance of the wetland ecosystems. The research also investigated the emergent biogeochemical-ecological similitude and scaling laws of wetland GHG fluxes by employing dimensional analysis from fluid mechanics. Two environmental regimes were found to govern the wetland GHG fluxes. The discovered similitude and scaling laws can guide the development of data-based mechanistic models to robustly predict wetland GHG fluxes under a changing climate and environment.

TABLE OF CONTENTS

CHAPTER	PAGE
Chapter 1: Introduction	1
1.1 Background and motivation	1
1.2 Research goal, hypotheses, and objectives	3
1.3 Organization of the dissertation	4
References	8
Chapter 2: Scaling-based robust empirical modeling of growing season net ecosystem exchange: application for deciduous forests	9
Abstract	9
2.1 Introduction	10
2.2 Materials and methods	15
2.2.1 Scaling concept	15
2.2.2 Steps of model development and application	19
2.2.3 Selection of optimal number of harmonics	20
2.2.4 Measurement of model sensitivity and uncertainty	21
2.2.5 Case study sites and data sets	22
2.3 Results and discussion	26
2.3.1 Optimal number of harmonics	26
2.3.2 Optimal reference time	26
2.3.3 Model evaluations for individual deciduous forests	28
2.3.4 Temporally robust predictions from model calibrations and validations	35
2.3.5 Spatiotemporal robustness: upscaling to a generalized deciduous forest model	37
2.3.6 Comparison of the model with the existing gap-filling techniques	40
2.3.7 Model sensitivity and uncertainty	42
2.3.8 Gap-filling of NEE time-series and predictions in relatively ungauged forests	44
2.3.9 Model generalization and applicability in different ecosystems	45
2.3.10 Data reliability and model limitations	48
2.4 Conclusions	50
References	52
Appendix	58
Chapter 3: Harmonic modeling of wetland net ecosystem exchange (NEE): Application of scaling for similarity and robust predictions	62
Abstract	62
3.1 Introduction	63
3.2 Materials and methods	68
3.2.1 The scaling-based harmonic model theory	68
3.2.2 Estimation of ESHA model parameters	69
3.2.3 The ESHA modeling algorithm for wetland NEE	70
3.2.4 Study wetlands and data sets	71
3.3 Results	74
3.3.1 Selection of an optimal model	74
3.3.2 Robust estimation of model parameters	77
3.3.3 Model evaluations and robust predictions	80

3.3.4	Model sensitivity and uncertainty	84
3.4	Discussion	87
3.4.1	Success in scaling and potential of finding wetland ecological similitudes	87
3.4.2	Gap-filing of wetland NEE time series and predictions.....	88
3.4.3	Applicability in different ecosystems and spatiotemporally robust predictions	89
3.4.4	Limitations of model evaluations.....	89
3.5	Conclusions.....	91
	References.....	93

Chapter 4: Relative linkages of canopy-level CO ₂ fluxes with the climatic and environmental variables for US deciduous forests		98
4.1	Introduction.....	99
4.2	Materials and methods	101
4.2.1	Study sites	101
4.2.2	Data sets.....	103
4.2.3	Data analysis and empirical modeling	106
4.3	Results.....	110
4.3.1	Correlation structure of the data.....	110
4.3.2	Dominant groups and orientation of the variables	111
4.3.3	Significant hidden factors	115
4.3.4	Relative carbon flux linkages of the climate and environmental variables.....	117
4.4	Discussion.....	125
4.4.1	Linking vertical carbon fluxes with the climatic and environmental variables	125
4.4.2	Data quality and uncertainty	127
4.4.3	Explanatory modeling and analysis	128
4.5	Conclusions.....	130
	References.....	132

Chapter 5: Data-analytics for comparative linking of net ecosystem exchange (NEE) with the climatic and ecohydrological variables for diverse ecosystems		138
	Abstract.....	138
5.1	Introduction.....	139
5.2	Materials and methods	143
5.2.1	Diverse ecosystem case study sites.....	143
5.2.2	The AmeriFlux data sets	146
5.2.3	The data-analytics framework.....	149
5.3	Results.....	153
5.3.1	Linear correspondences of NEE, climatic and ecohydrological variables.....	153
5.3.2	Relative orientations and groupings of NEE and driving variables	154
5.3.3	Dominant biophysical variables based on optimal, orthogonal latent factors.....	157
5.3.4	Climatic and environmental process components driving NEE	159
5.3.5	Quantitative relative linking of NEE through explanatory PLSR.....	160
5.4	Discussion	166
5.4.1	On the variable groupings and their comparative linkages	166
5.4.2	On the effectiveness of the data-analytics method.....	169
5.4.3	On the reliability and uncertainty of empirical data sets.....	172
5.5	Conclusions.....	172
	References.....	175

Chapter 6: Relative environmental controls and empirical predictive modeling of daytime CO ₂ fluxes of coastal salt marshes	184
6.1 Introduction.....	185
6.2 Materials and methods	188
6.2.1 Study sites and data set	188
6.2.2 Data preparation and data statistics.....	193
6.2.3 Data Analysis and predictive modeling	194
6.2.4 Measures of modeling performance.....	197
6.3 Results.....	198
6.3.1 Correlation analysis	198
6.3.2 Relative orientations and groupings of the variables	198
6.3.3 Extracted orthogonal latent factors	201
6.3.4 Partial least squares modeling in non-linear domain	203
6.3.5 Predictive model development.....	205
6.4. Discussion.....	208
6.4.1 F _{CO2} linkages of the climatic and environmental variables.....	208
6.4.2 Predictive model fitting with a minimum parameter set.....	213
6.4.3 Significance of the estimated parameters.....	214
6.4.4 Generalization of the model.....	216
6.5 Conclusions.....	217
References.....	218
Chapter 7: Data-driven mechanistic modeling of CH ₄ emission from coastal wetlands:	
Application of a data-analytics framework.....	224
Abstract.....	224
7.1 Introduction.....	225
7.2 Materials and methods	227
7.2.1 Study sites and data set	227
7.2.2 Data preparation and data statistics.....	228
7.2.3 Data analysis and predictive modeling	230
7.3 Results.....	232
7.3.1 Correlation analysis	232
7.3.3 Extracted orthogonal latent factors	234
7.3.4 Partial least squares modeling in non-linear domain	236
7.3.5 Predictive model development.....	239
7.4 Discussion.....	243
7.4.1 Relative CH ₄ flux linkages of the participatory physical drivers	243
7.4.2 Parsimonious predictive modeling.....	248
7.4.3 Significance of the estimated parameters.....	250
7.5 Conclusions.....	253
References.....	254
Chapter 8: Investigation of biogeochemical similitudes, environmental regimes, and scaling relationships of greenhouse gas (GHG) fluxes from coastal wetlands	257
Abstract.....	257
8.1 Introduction.....	258
8.2 Materials and methods	261
8.2.1 Study sites and data set	261
8.2.2 Data preparation and data statistics.....	262
8.2.3 Analysis methodology	263

8.3	Results.....	266
8.3.1	Formulation of dimensionless numbers	266
8.3.2	Identification of important and meaningful dimensionless numbers	271
8.3.3	Examination of similitude and environmental regimes	276
8.3.4	Development of scaling laws and modelling	283
8.4	Discussion.....	287
8.4.1	Biogeochemical-ecological similitudes and environmental regimes	287
8.4.2	Scaling laws and modeling	291
8.5	Conclusions.....	293
	References.....	295
	Chapter 9: Summary and conclusions.....	297
	VITA.....	302

LIST OF TABLES

TABLE		PAGE
2.1	Summary of the selected case study deciduous forests and the associated growing season (June to September) hourly NEE data periods.....	17
2.2	Growing season (June to September) averages of NEE and associated environmental/meteorological variables for the calibration years for the study sites.....	25
2.3	Temporal ensembles (multi-year site-specific averages over May-October) and spatiotemporal means of daily estimated model parameters for the five deciduous forests.....	28
2.4	Results of calibrations and validations for the individual forest-specific NEE models.....	35
2.5	Performance of the generalized deciduous forest model in predicting the hourly NEE of different days and years for the five study forests.....	39
2.6	Comparison of the daytime and nighttime predicted means against the observed daytime and nighttime means of NEE for the calibration, validation and generalized model for all the study forests	40
3.1	Summary of the study sites and NEE data collection period.....	71
3.2	Means of (averaged over the calibration period) estimated model parameters at study sites.....	76
3.3	Summary of the results for site specific model calibration and validation at different study sites.....	80
4.1	Means and standard deviations (in parenthesis) of the observed carbon fluxes, and climate and environmental variables for the eight broad-leaf deciduous forest sites of USA.....	102
4.2	Pearson correlation coefficients between the carbon fluxes (F_{CO_2}) and the climate/environmental variables for the eight deciduous forest study sites (Blank indicates missing data).....	110
4.3	Dominant latent factors extracted from the data matrices of the eight deciduous forest study sites.....	114
4.4	Coefficients (β) of the normalized (dimensionless) PLSR models of carbon fluxes (F_{CO_2}) for different deciduous forest study sites.	119

5.1	Bio-physical summary of the diverse ecosystem study sites from the AmeriFlux network	143
5.2	Summary of data records and statistics (averages and standard deviations) for the observed hourly NEE, climate and ecohydrological variables of the six study sites. Standard deviations are given in parenthesis.....	145
5.3	Pearson correlation coefficients between the NEE and climatic/ecohydrological variables for the diverse ecosystem study sites.	145
5.4	Dominant latent factors and optimized loadings of process variables for the diverse ecosystem study sites.....	156
5.5	Coefficients (BETA) of standardized (Z-score) PLSR models of NEE for different study ecosystems.....	161
5.6	Aggregated relative linkages of the major biophysical groups with NEE for different ecosystems.....	162
6.1	Data statistics (mean, standard deviation, maximum, and minimum) of the observed (76 observations) climatic, environmental, and biological variables for the study wetlands.....	189
6.2	Pearson correlation coefficient between the daytime carbon fluxes (F_{CO_2}) and the corresponding climatic and environmental variables for the study area for both $WD>0$ (water table above the soil marsh) and $WD<0$ (water table below the soil marsh) conditions.....	197
6.3	Extracted latent factors from the factor analysis for the study wetlands for both $WD>0$ and $WD<0$ (shaded area in the table) conditions.....	199
6.4	Coefficients (BETA) of the log-transformed, normalized (dimensionless) PLSR models (2 PLS component based) of the daytime F_{CO_2} for both $WD>0$ and $WD<0$ conditions. R^2 , RSR, and MAE refer, respectively, to coefficient of determination, ratio of root-mean-square error to the standard deviation of observations, and mean absolute error.....	202
6.5	The estimated model parameters (equation 2) of the predicted model (obtained from the mean of the parameters through 10000 iterations) of daytime F_{CO_2} for the Waquoit bay, MA salt marshes.....	205
7.1	Data statistics (means, standard deviations, maximum, and minimum) of the observed (66 observations) climatic, environmental, and biological variables for the study wetlands.....	227
7.2	Pearson correlation coefficients between the CH_4 emission fluxes (F_{CH_4}) and the corresponding climatic and environmental variables for the study area for both $WD>0$ (water table above the soil marsh) and $WD<0$ (water table below the soil marsh) conditions.....	230

7.3	Extracted latent factors from the factor analysis for the study wetlands for both WD>0 and WD<0 (shaded area in the table) conditions.....	232
7.4	Coefficients (BETA) of the log-transformed, normalized (dimensionless) 3-component (optimized) PLSR models of the F _{CH₄} for both WD>0 and WD<0 conditions	236
7.5	The estimated parameters of the predicted model (obtained from the mean of the parameters through 10000 iterations) of F _{CH₄} for the Waquoit bay, MA wetlands. The mean p-values of the estimated parameters are given in parenthesis.....	239
8.1	List of variables, units, and data range and mean for all the study wetlands.....	259
8.2	List of variables used for the dimensional analysis of wetland GHG fluxes.....	261
8.3	Unit conversion factors of all the variables used in dimensional analysis.....	262
8.4	Pearson correlation matrix for wetland CO ₂ sequestration flux dataset for all the study ponds, showing linear correlations among the participatory dimensionless numbers.....	269
8.5	Latent factors extracted from the data matrix for F _{CO₂} . Bold indicates variables that loaded highly on each factor.....	270
8.6	Pearson correlation matrix for wetland CH ₄ flux dataset for all the study ponds, showing linear correlations among the participatory dimensionless numbers.....	271
8.7	Latent factors extracted from the data matrix for F _{CH₄} . Bold indicates variables that loaded highly on each factor.....	272
8.8	The estimated model parameters and model fitting efficacy for different modeling scenarios to predict CO ₂ dimensionless number (Shaded area is for WD<0 condition).....	281
8.9	The estimated model parameters and model fitting efficiencies for different modeling scenarios to predict CH ₄ dimensionless numbers.....	285
8.10	Mean based comparison of the physical factors of CO ₂ and CH ₄ fluxes for low (SL and HP site) and high environmental regimes (GP and EP site) based on the observations.....	288

LIST OF FIGURES

FIGURE		PAGE
1.1	Chart showing the research hypotheses and organization of the dissertation	5
2.1	Schematics showing (a) comparison of the 24 hour diurnal cycles of NEE for 61 th (representing dormant season) and 158 th (representing growing season) Julian day of Morgan forest to demonstrate the significant temporal disparity in the shape of the NEE curve and (b) the time averaged growing season (June to September) NEE having well defined diurnal cycles for five different deciduous forests considered in the study.....	16
2.2	Schematic of the conceptual framework for development and application of the scaled net ecosystem exchange (NEE) model.....	18
2.3	Study area and geographic locations of the selected AmeriFlux deciduous forest sites.....	24
2.4	Plots of the normalized AIC_{NEE^*} with the number of harmonics (W) for different study forests.....	26
2.5	Impact of choosing different reference time (t_{ref}) on the modeling efficiency, R^2 for the five study forests.....	27
2.6	Daily time-series of estimated parameters during the four years of calibration for the five deciduous forests, demonstrating the spatiotemporal robustness of the model parameters. The a_0 , a_1 , and b_1 , respectively, refer to \hat{a}_0 , \hat{a}_1 and \hat{b}_1 where the hat (^) represents estimations.....	30
2.7	Plots of the daily estimates of the principal model parameter, \hat{a}_0 with the corresponding (as normalized by the respective site-specific averages) (a) photosynthetic active radiation ($PAR_{daily}/PAR_{average}$), (b) air temperature ($TA_{daily}/TA_{average}$), and (c) soil water content ($SWC_{daily}/SWC_{average}$) over the four different years of model calibrations for different sites.....	31

2.8	Plots of the daily estimates of the principal model parameter, \hat{a}_1 with the corresponding (as normalized by the respective site-specific averages) (a) photosynthetic active radiation ($PAR_{daily}/PAR_{average}$), (b) air temperature ($TA_{daily}/TA_{average}$), and (c) soil water content ($SWC_{daily}/SWC_{average}$) over the four different years of model calibrations for the study sites.....	32
2.9	Plots of the daily estimates of the principal model parameter, \hat{b}_1 with the corresponding (as normalized by the respective site-specific averages) (a) photosynthetic active radiation ($PAR_{daily}/PAR_{average}$), (b) air temperature ($TA_{daily}/TA_{average}$), and (c) soil water content ($SWC_{daily}/SWC_{average}$) over the four different years of model calibrations for different sites.....	33
2.10	Plots of the hourly predicted and observed NEE of different years for the (a) Harvard Forest, (b) Missouri Ozark, (c) Morgan State Forest, (d) UMBS, and (e) Willow Creek Forest using the generalized deciduous forest model. Dashed line represents the 1:1 perfect fit line.....	38
2.11	Plot of dimensionless parameter sensitivity coefficients ($S_{\hat{a}_0}, S_{\hat{a}_1}, S_{\hat{b}_1}, S_{\hat{a}_2}$, and $S_{\hat{b}_2}$) with the hours of any diurnal cycle, indicating the model sensitivity for the (a) Harvard Forest, (b) Missouri Ozark and Willow Creek, (c) Morgan State Forest, and UMBS Forest ($S_{a0}, S_{a1}, S_{a2}, S_{b1}$, and S_{b2} , respectively, refer to model sensitivity to $\hat{a}_0, \hat{a}_1, \hat{a}_2, \hat{b}_1$ and \hat{b}_2	43
2.12	Standard deviations, estimated from eqn. 6, of the predicted hourly NEE as an overall measure of the model uncertainty for simultaneously changing all parameters by their respective standard deviations at the five deciduous forests.....	45
3.1	Study area and geographic locations of the selected AmeriFlux wetland study sites.....	70
3.2	Plot of the hourly net ecosystem exchange (NEE) of Taylor Slough and Lost Creek sited for 2008 showing variation in diurnal cycles of NEE (upper panel). 100 th and 200 th Julian days are zoomed in for Taylor slough and Lost creek from 2008 hourly NEE data to see the diurnal variation of NEE for these two sites at different season (lower panel).....	73
3.3	Optimization of the ESHA model for the selection of (a) maximum number of harmonics required for model simulation by plotting the normalized AIC_{NEE^*} with the number of harmonics (W) and (b) best reference time (t_{ref}) from the 24-hour period that could be used for modeling from a plot	74

	of coefficient of determination, R^2 and different diurnal hours for two study wetlands.....	
3.4	Daily time-series of estimated parameters during the four years of calibration for the for the study wetlands, demonstrating the temporal robustness of the model parameters. The a_0 , a_1 , and b_1 , respectively, refer to \hat{a}_0 , \hat{a}_1 and \hat{b}_1 where the hat (^) represents estimations.....	77
3.5	Plots of the daily estimates of the principal model parameter, \hat{a}_0 and 1 st harmonic parameters \hat{a}_1 and \hat{b}_1 with the corresponding (as normalized by the respective site-specific averages), air temperature ($TA_{\text{daily}}/TA_{\text{average}}$), wind speed ($WS_{\text{daily}}/WS_{\text{average}}$) and photosynthetic active radiation ($PAR_{\text{daily}}/PAR_{\text{average}}$) for the year 2008 at two wetland ecosystems. . The a_0 , a_1 , and b_1 , respectively, refer to \hat{a}_0 , \hat{a}_1 and \hat{b}_1 where the hat (^) represents estimations.....	79
3.6	Calibration plots of the hourly predicted and observed NEE of different calibration periods for the (a) Taylor Slough (US-Esm) and, (b) Lost Creek (US-Los) using temporally ensemble means of site-specific model parameters. Dashed red line represents the 1:1 perfect fit line.....	80
3.7	Time series plot of the hourly predicted and observed NEE of different validation period for the Taylor Slough (US-Esm) and Lost Creek (US-Los) sites showing impressive validation.....	82
3.8	Variation of dimensionless parameter sensitivity coefficients ($S_{\hat{a}_0}, S_{\hat{a}_1}, S_{\hat{b}_1}, S_{\hat{a}_2}$, and $S_{\hat{b}_2}$) with the hours of any diurnal cycle, indicating the model sensitivity for the study wetlands. Since same 12:00 PM was used as a reference time (t_{ref}) for both the wetlands, of dimensionless parameter sensitivity coefficients are identical. ($S_{a0}, S_{a1}, S_{a2}, S_{b1}$, and S_{b2} , respectively, refer to model sensitivity to $\hat{a}_0, \hat{a}_1, \hat{a}_2, \hat{b}_1$ and \hat{b}_2) (upper panel). Bottom panel figure showed standard deviations of the predicted hourly NEE as an overall measure of the model uncertainty for simultaneously changing all parameters by their respective standard deviations at two wetland sites.....	84
4.1	Locations of the selected eight deciduous forest study sites.....	100
4.2	The data-analytics methodology to determine the relative carbon flux linkages of different climate and environmental variables.....	105
4.3	Biplots obtained from principal component analysis showing the groupings and interrelation patterns of the climate, environmental, and biological	110

	variables for (a) Bartlett Experimental Forest (NH), (b) Harvard Forest (MA), (c) UMBS Forest (MI), and (d) Missouri Ozark Forest (MO). Percent variance explained by each PC is shown in parenthesis.....	
4.4	Biplots obtained from principal component analysis showing the groupings and interrelation patterns of the climate, environmental, and biological variables for (a) Silas Little Forest (NJ), (b) Willow Creek Forest (WI), (c) Morgan Monroe State Forest (IN), and (d) Ohio Oak Openings Forest (OH). Percent variance explained by each PC is shown in parenthesis.....	111
4.5	Plot of cross-validated (a) normalized AIC and (b) fitting efficiency (R^2) for F_{CO_2} with the number of incorporated partial least squares (PLS) components.	116
4.6	Variable importance on the partial least squares projection (VIP) of different predictors for the response variable (F_{CO_2}) for (a) Bartlett Experimental Forest (NH), (b) Harvard Forest (MA), (c) UMBS Forest (MI), (d) Missouri Ozark Forest (MO), (e) Silas Little Forest (NJ), (f) Willow Creek Forest (WI), (g) Morgan Monroe State Forest (IN), and (h) Ohio Oak Openings Forest (OH). Dashed line indicates predictors having VIP score greater than unity (1.0).....	118
5.1	Locations of six diverse AmeriFlux ecosystem study sites.....	142
5.2	The data-analytics framework.	148
5.3	Biplots from principal component analysis, showing the interrelation and grouping patterns of climatic, ecohydrological and biological variables for (a) Fermi Prairie (IL), (b) Sylvania Wilderness (MI), (c) Niwot Ridge Forest (CO), (d) Everglades (FL), (e) Mead Irrigated (NE), and (f) Morgan Monroe State Forest (IN). Percent variance explained by each PC is shown in parentheses.....	153
5.4	Plot of cross-validated (a) normalized AIC and (b) fitting efficiency (R^2) of NEE with the number of partial least squares components.....	159
5.5	Variable importance on the partial least squares projection (VIP) scores for different predictors of NEE for (a) Fermi Prairie (IL), (b) Sylvania Wilderness (MI), (c) Niwot Ridge Forest (CO), (d) Everglades (FL), (e) Mead Irrigated (NE), and (f) Morgan Monroe State Forest (IN). Dashed line indicates variables with VIP score greater than unity (1.0).....	160
6.1	Locations of the four wetland sites in the Waquoit Bay, MA.....	187
6.2	Measured CO_2 fluxes for the different days and diurnal hours of 2013 at the four study ponds of Waquoit Bay, MA. Positive values of the fluxes indicate downward (atmosphere to soil) fluxes (i.e., potential sequestration).	188

6.3	The data-analytics methodology to determine the relative linkages of different climatic and environmental variables with FCO ₂ for the Waquoit Bay salt marshes.....	193
6.4	Biplots obtained from the loadings of the 1 st two principal components (PC) showing the groupings and interrelation patterns of the climatic, environmental, and biological variables for both (a) WD>0 and (b) WD<0 conditions. Percent variance explained by each PC is shown in parenthesis. PAR, RD, AT, ST, pH, SS, SM, and F _{CO₂} refer, respectively, to photosynthetic active radiation, water depth relative to water marsh surface, air temperature, soil temperature, soil pH, soil salinity, soil moisture, and daytime CO ₂ exchanges of the salt marshes.....	198
6.5	Plot of 10-fold cross-validated (a) fitting efficiency, R ² and (b) Normalized AIC for both WD>0 and WD<0 conditions of instantaneous F _{CO₂} with the number of incorporated partial least squares (PLS) components instantaneous.....	201
6.6	Plot shows the histograms of the estimated parameters of the prediction model (option 1: without soil moisture) obtained from the Monte Carlo cross validation through 10000 iterations.....	204
6.7	Plot of the predicted and observed instantaneous daytime F _{CO₂} of 2013 growing season for the four study ponds of Waquoit Bay, MA. The dashed line represents the 1:1 perfect fit line.....	205
6.8	Change of daytime CO ₂ exchange (F _{CO₂}) with the change of soil salinity for all the four study ponds. The slope of linear fit indicated a negative correlation between them.....	210
7.1	Measured CH ₄ emission fluxes for the different days and diurnal hours of 2013 at the four study ponds of Waquoit Bay, MA. The negative values of the fluxes indicate emission from the marsh surface.....	225
7.2	Biplots obtained from the loadings of the 1 st two principal components (PC) showing the groupings and interrelation patterns of the climatic, environmental, and biological variables for both (a) WD>0 and (b) WD<0 conditions. Percent variance explained by each PC is shown in parenthesis. PAR, WD, AT, ST, pH, SS, SM, and F _{CH₄} refer, respectively, to photosynthetic active radiation, water depth relative to water marsh surface, air temperature, soil temperature, soil pH, soil salinity, soil moisture, and CH ₄ from the salt marshes.....	231

7.3	Plot of 10-fold cross-validated (a) fitting efficiency, R^2 and (b) Normalized AIC for both $WD > 0$ and $WD < 0$ conditions of instantaneous F_{CH_4} with the number of incorporated partial least squares (PLS) components.....	235
7.4	Plot shows the histogram of the estimated parameters of the predictive model for $WD > 0$ condition obtained from the Monte Carlo cross validation through 10000 iterations.....	240
7.5	Plot shows the histogram of the estimated parameters of the 4- parameter predictive model for $WD < 0$ condition obtained from the Monte Carlo cross validation through 10000 iterations.....	241
7.6	Plot shows the histogram of the estimated parameters of the 2-parameter predictive model obtained from the Monte Carlo cross validation through 10000 iterations.	242
7.7	Plot of the observed versus predicted CH_4 fluxes of the coastal wetland for (a) $WD > 0$ condition, (b) $WD < 0$ condition with three variable model, and (C) $WD < 0$ condition for the one variable (WD only) model.....	242
8.1	PCA biplots showing correlation strength and orientation of variables for wetland F_{CO_2} for all the study ponds of Waquoit Bay, MA. Percent variance explained by each PC is shown in parentheses. R_1 and R_2 are carbon flux group ($R_1 = [(F_{CO_2} * t) / (WD * SS)]$ and $R_2 = [(F_{CO_2} * WD^2) / (PAR * t)]$); P_1 and P_4 are the radiation-redox group ($P_1 = [(PAR * t^3) / (WD^3 * SS)]$ and $P_4 = [(PAR * t^3) / (WD^3 * H^+)]$); P_2 is the hydro-climatic group $[(ST * C_p * t^2) / (WD^2)]$; and P_3 is the salinity-redox group $[SS / H^+]$. F_{CO_2} , PAR , WD , ST , SS , C_p , H^+ , and t refer, respectively, to net CO_2 sequestration flux, photosynthetic active radiation, water depth relative to soil surface, soil temperature, soil salinity, specific heat at constant pressure, hydrogen ion concentration, and time.	269
8.2	PCA biplots showing correlation strength and orientation of variables for wetland F_{CH_4} for all the study ponds of Waquoit Bay, MA. Percent variance explained by each PC is shown in parentheses. R_1 and R_2 are carbon flux group ($R_1 = [(F_{CH_4} * t) / (WD * SS)]$ and $R_2 = [(F_{CH_4} * WD^2) / (PAR * t)]$); P_1 and P_4 are the radiation-redox group ($P_1 = [(PAR * t^3) / (WD^3 * SS)]$ and $P_4 = [(PAR * t^3) / (WD^3 * H^+)]$); P_2 is the hydro-climatic group $[(ST * C_p * t^2) / (WD^2)]$; and P_3 is the salinity-redox group $[SS / H^+]$. F_{CH_4} , PAR , WD , ST , SS , C_p , H^+ , and t refer, respectively, to CH_4 emission flux, photosynthetic active radiation, water depth relative to soil surface, soil temperature, soil salinity, specific heat at constant pressure, hydrogen ion concentration, and time.	272
8.3	Plot of the predictor dimensionless groups ($\log_{10} P_1$ and $\log_{10} P_2$) with the response dimensionless group ($\log_{10} R_1$), revealing collapse of different variables on the power-law scaling relationships. R_1 is the CO_2 flux group	274

$[(F_{CO_2} * t) / (WD * SS)]$; P_1 is the radiation-redox group $[(PAR * t^3) / (WD^3 * SS)]$; and P_2 is the hydro-climate group $[(ST * C_p * t^2) / (WD^2)]$

- 8.4 Plot of the predictor dimensionless groups ($\log_{10}P_1$ and $\log_{10}P_2$) with the response dimensionless group ($\log_{10}R_1$) for two different hydrological regimes (flood and no-flood), revealing collapse of different variables on the power-law scaling relationships. R_1 is the CO_2 flux group $[(F_{CO_2} * t) / (WD * SS)]$; P_1 is the radiation-redox group $[(PAR * t^3) / (WD^3 * SS)]$; and P_2 is the hydro-climate group $[(ST * C_p * t^2) / (WD^2)]$ 275
- 8.5 Plot of the predictor dimensionless groups ($\log_{10}P_1$ and $\log_{10}P_2$) with the response dimensionless groups ($\log_{10}R_1$ and $\log_{10}R_2$), revealing collapse of different variables on the power-law scaling relationships. R_1 and R_2 are the CH_4 flux group ($R_1 = [(F_{CH_4} * t) / (WD * SS)]$ and $R_2 = [(F_{CH_4} * WD^2) / (PAR * t)]$); P_1 is the radiation-redox group $[(PAR * t^3) / (WD^3 * SS)]$; and P_2 is the hydro-climate group $[(ST * C_p * t^2) / (WD^2)]$ 278
- 8.6 Plot of the predictor dimensionless groups ($\log_{10}P_1$ and $\log_{10}P_2$) with the response dimensionless groups ($\log_{10}R_1$ and $\log_{10}R_2$) for two different hydrological regimes (flood and no-flood), revealing collapse of different variables on the power-law scaling relationships. R_1 and R_2 are the CH_4 flux group ($R_1 = [(F_{CH_4} * t) / (WD * SS)]$ and $R_2 = [(F_{CH_4} * WD^2) / (PAR * t)]$); P_1 is the radiation-redox group $[(PAR * t^3) / (WD^3 * SS)]$; and P_2 is the hydro-climate group $[(ST * C_p * t^2) / (WD^2)]$ 279

Chapter 1: Introduction

1.1 Background and motivation

The land-atmospheric exchanges of carbon dioxide (CO₂) and methane (CH₄) are major drivers of global warming and climatic changes. The annual escalation rate of atmospheric CO₂ concentration is 1.4 ppm per year (1960-2005), and the rate is increasing (1.9 ppm per year based on average during 1995-2005) (IPCC, 2007). The atmospheric concentration of CH₄ has increased over 150% during 1750 to 2011 (Hartmann et al., 2013; Zhu et al., 2015). The global warming potential of CH₄ (compared to CO₂) is 34 and 86 for, respectively, 100 and 20 years (Stocker et al., 2003). Furthermore, the lifespan of CH₄ in the atmosphere is very high compared to that of CO₂; CH₄ is one of the contributors of the ozone layer depletion in the stratosphere (Randeniya et al., 2002). The greenhouse gas (GHG) fluxes also indicate the dynamics and potential storage of carbon in terrestrial and wetland ecosystems. Appropriate modeling and prediction tools can provide valuable insights into the ecosystem carbon dynamics, and aid in the development of engineering and management strategies to limit emissions of GHGs and enhance carbon sequestration. Therefore, explanatory and predictive models of ecosystem GHG fluxes are essential ecological engineering tools for ecosystem managers to set priorities for ecosystem conservation and adaptation; to achieve sustainability and resilience under a changing climate and environment.

Turbulent land-atmospheric fluxes of vertical CO₂ contribute to ecosystem-scale carbon budget by balancing between photosynthesis and respiration (both autotrophic and heterotrophic) (Heimann and Reichstein 2008). If below-canopy carbon storage is considered with turbulent carbon exchange, the flux is then called net ecosystem exchange (NEE) (Xiao et al., 2008). NEE follows periodic oscillations, whereas available measurements are often not continuous due to instrumental failure, inclement weather condition, and probable breach in micrometeorological

hypothesis (Moffat et al., 2007). Therefore, a standard method is necessary to generate fine-resolution, continuous carbon fluxes from a limited number of observations for a proper estimation of the net ecosystem carbon budget.

The carbon flux-climate feedback process is sensitive to an array of environmental and ecological variables at different spatiotemporal scales. Although much research has been conducted to characterize the role of key environmental stressors and climatic variability on land-atmospheric carbon and heat fluxes (e.g., Schmidt et al. 2011), their robust linkage patterns are yet to be understood well in variable space and time. Investigation of the relative linkages of climatic and environmental variables with the GHG fluxes is also a major step towards building relatively simple empirical to complex mechanistic models for predictions.

Coastal wetland illustrated highest sequestration rate of carbon among all ecosystems (Chmura et al., 2003; Duarte et al., 2005). The carbon storage and sequestration capacity of coastal and marine wetlands are often referred to as “blue carbon” (in contrast to forest or “green carbon”). Blue carbon is an emerging concept for coastal management that can potentially attract much private and public investment in coastal protection and restoration (Nellemann et al. 2009). Mechanistic knowledge on the relative linkages of the wetland GHG fluxes with the biogeochemical and ecological drivers would guide the development of parsimonious (i.e., involving minimum parameters) predictive models that are simple, computationally inexpensive and handy for the end-users and decision makers. Such modeling tools can provide proper estimations and predictions of net ecosystem carbon budget (NECB) for wetlands on an appropriate scale (e.g., annual) to develop GHG offset protocols.

Furthermore, available mechanistic and empirical models of ecosystem GHG fluxes are mostly site-specific and can rarely predict the GHG fluxes that are robust at variable time, space and process scales. Investigations and utilization of scaling and similitude laws of the ecosystem

GHG fluxes can lead to the identification of crucial environmental regimes, guiding the development of scale-invariant models for robust predictions of ecosystem GHG fluxes and carbon storage. This dissertation focuses on the application of data-analytics to develop engineering models and relatively simple tools by employing a range of empirical and semi-mechanistic approaches to robustly predict ecosystem GHG fluxes at variable scales.

1.2 Research goal, hypotheses, and objectives

The overall research goal of this dissertation is to *unravel the organizing principles of ecosystem GHG fluxes to develop empirical and data-based mechanistic models for obtaining robust predictions in diverse terrestrial and wetland ecosystems*. The research is expected to provide crucial understanding and insights into the linkage and emergence patterns of ecosystem GHG fluxes. The predictive modeling tools for CO₂ and CH₄ fluxes are expected to assist ecologists, engineers and stakeholders with a proper management of ecosystem carbon and associated services under a changing climate and environment.

The entire dissertation is designed on the evaluation of three research hypotheses and four supporting objectives. The associated research hypotheses are as follows:

Hypothesis 1: Scaling-based empirical modeling provides spatiotemporal robust predictions of the different diurnal cycles of terrestrial and wetland GHG fluxes.

Hypothesis 2: Data-analytics can lead to the development of parsimonious data-based mechanistic models of ecosystem GHG fluxes.

Hypothesis 3: Wetland GHG fluxes follow emergent similitude, environmental regimes, and scaling laws.

The hypotheses are tested based on the following specific research objectives:

1. Develop scaling and similarity based empirical models for spatiotemporally robust predictions of the diurnal cycles of terrestrial and wetland NEE.

2. Develop a systemic data-analytics framework to determine the latent interrelation and relative linkage patterns of different climatic and environmental variables with the terrestrial CO₂ fluxes across different ecosystems.
3. Determine the key drivers of wetlands GHG fluxes using the data-analytics framework, and develop data-based mechanistic models to predict GHG fluxes of tidal wetlands.
4. Unravel the emergent similitudes, environmental regimes, and scaling laws associated with the wetland GHG fluxes.

1.3 Organization of the dissertation

The dissertation is organized into 9 chapters (Figure 1.1). Chapter 1 displays the introduction and overview of the dissertation, including background and motivation, research goal, hypothesis and the supporting objectives.

Chapter 2 and 3 test the first hypothesis by pursuing objective 1. Chapter 2 demonstrates the development and application of an extended stochastic harmonic algorithm (ESHA) framework for a robust estimation and prediction of hourly NEE from different forest ecosystems. The model was calibrated and validated for the growing season (June-September) diurnal cycles of NEE during 2005-2013 for five AmeriFLUX deciduous forest sites. The site-specific models were successfully scaled up to a generalized deciduous forest model to robustly predict hourly NEE for different days across different ecosystems. The chapter is submitted to *Agriculture and Forest Meteorology* for publication with the title, “Scaling-based robust empirical modeling of growing season net ecosystem exchange: Application for deciduous forests” (in revision).

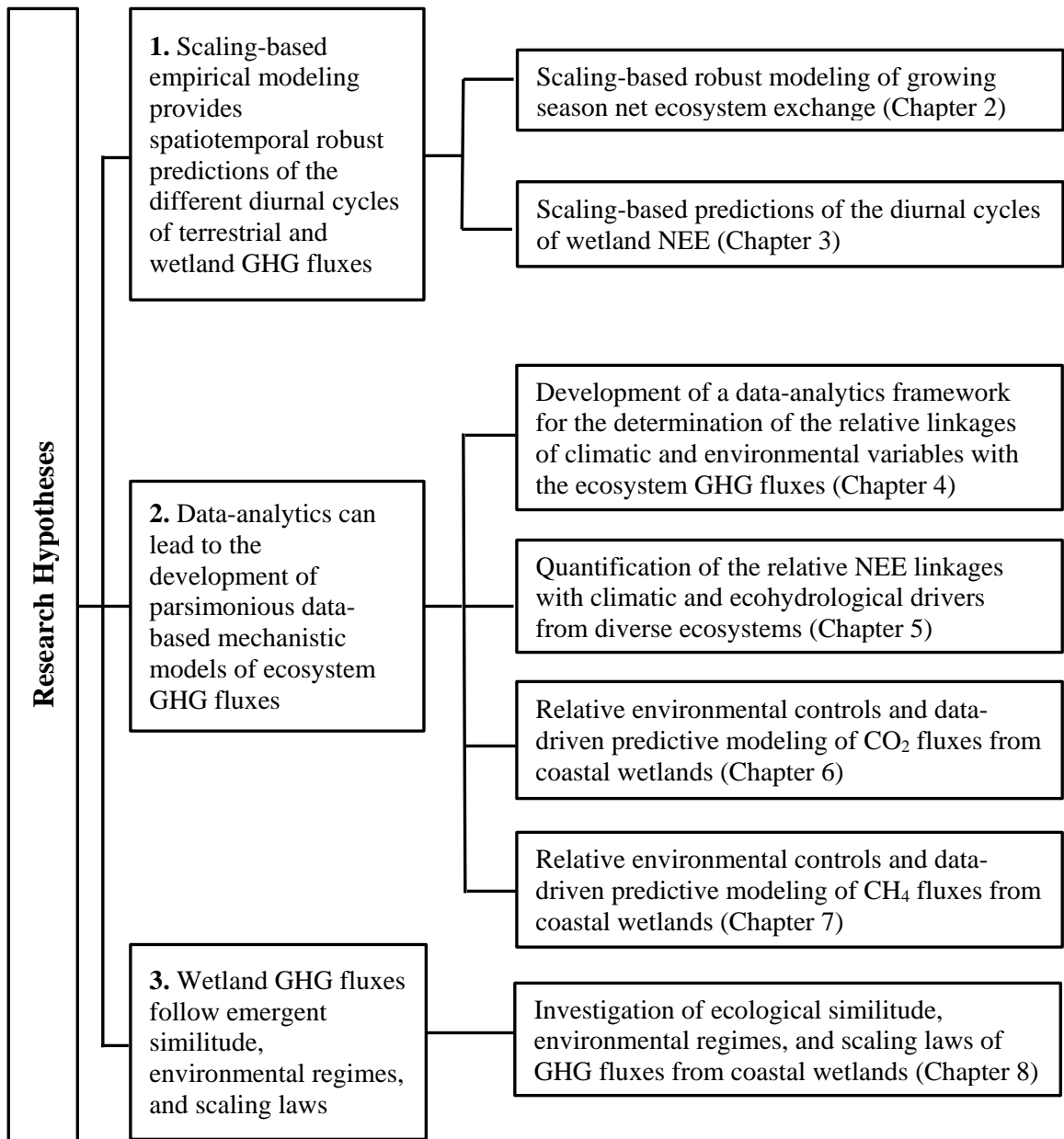


Figure 1.1: Chart showing the research hypotheses and organization of the dissertation

Chapter 3 presents a scaling-based harmonic model for robust predictions and simulations of the diurnal cycle of wetland NEE from a single reference observation. The harmonic model utilizes the ESHA algorithm presented in chapter 2. One inland wetland and one coastal wetland from the AmeriFlux network were the case study sites, and the developed models were calibrated and validated to predict different diurnal cycles of NEE by using a temporally ensemble set of model parameters for the respective study sites. Sensitivity and uncertainty of the estimated parameters were also computed to demonstrate model robustness.

Chapter 4-7 test the second hypothesis by pursuing objectives 2 and 3. Chapter 4 introduces a new, systematic data-analytics framework to determine the interrelation patterns, similarity based groupings, and relative linkages of different climatic and environmental variables with the ecosystem GHG fluxes. The data-analytics approach was tested by quantitatively linking the canopy-level, half-hourly CO₂ fluxes with the different process variables of eight deciduous forests from the AmeriFlux network. The chapter is published with the title, “Relative Linkages of Canopy-level CO₂ Fluxes with the Climatic and Environmental Variables for US Deciduous Forests” in *Environmental Management*, 55-4(943-960), 2015.

Chapter 5 tests the utility and applicability of the data-analytics framework in diverse ecosystems. The relative orientations, linkages and similarity-based grouping patterns of different climatic and environmental variables with hourly NEE were quantified for six diverse ecosystems, including wetlands, grasslands, croplands, deciduous broadleaf forest, evergreen needle-leaf forest, and mixed forest.

Leveraging the data-analytics framework, Chapter 6 presents a parsimonious, non-linear (power-law), data-based mechanistic model to predict CO₂ fluxes of coastal wetlands. The most significant drivers and predictors of the wetland CO₂ fluxes were identified by determining the relative control and linkages of different abiotic drivers of CO₂ fluxes based on the data-

analytics. Four coastal salt marsh sites of Waquoit Bay, MA were considered as the case studies, representing a moderate ecological and biogeochemical gradient.

Chapter 7 demonstrates the development of a parsimonious, non-linear (power-law), data-based mechanistic model to predict CH₄ emissions from coastal wetlands by using the observations of Waquoit Bay study sites. The model input variables were selected based on the mechanistic attributes of the most important drivers of CH₄ emissions. The model is capable of predicting the wetland CH₄ emissions under a range of salinity, temperature and inundation conditions. Models presented in chapter 6 and 7 can jointly estimate the net ecosystem carbon balance (NECB) (i.e., potential sequestration) of the coastal salt marshes.

Chapter 8 tests the third hypothesis in pursuit of objective 4. It presents the investigation of biogeochemical-ecological similitude, environmental regimes, and emerging scaling laws and organizing principles for GHG fluxes of coastal wetlands. The most important ecological drivers of CO₂ and CH₄ fluxes based on analytics of the Waquoit Bay data were used for developing the scaling relationships. Dimensional scaling models of the GHG fluxes were also presented for the coastal wetlands.

Chapter 9 summarizes the overall findings of the dissertation research, and discusses the utility of the research findings and tools to the user-communities for managing ecosystem carbon and mitigating climatic changes. The chapter also provides recommendations for future research to further the scientific knowledge and understanding of ecosystems' biological fluxes; and to enhance the modeling and engineering tools for robust estimations and predictions of ecosystem carbon based on emergent scaling laws and organizing principles.

References

- Chmura, G.L., Anisfeld, S.C., Cahoon, D.R., Lynch, J.C., 2003. Global carbon sequestration in tidal, saline wetland soils. *Global Biogeochemical Cycles*, 17 (4), 1111.
- Duarte, C.M., Middelburg, J., Caraco, N., 2005. Major role of marine vegetation on the oceanic carbon cycle. *Biogeosciences*, 2 (1), 1–8.
- Hartmann, D. L., Klein Tank, A. M. G., Rusicucci, M., Alexander, L. V., Broenniman, B., Charabi, Y., ... Kent, E., 2013. Observations: atmosphere and surface.
- Heimann, M., Reichstein, M., 2008. Terrestrial ecosystem carbon dynamics and climate feedbacks. *Nature*, 451(7176), 289-292.
- IPCC., 2007. Climate change 2007: The physical science basis. *Agenda*, 6(07), 333.
- Moffat, A. M., Papale, D., Reichstein, M., Hollinger, D. Y., Richardson, A. D., Barr, A. G., ... Stauch, V. J., 2007. Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. *Agricultural and Forest Meteorology*, 147(3), 209-232.
- Nellemann, C., Corcoran, E., (Eds.) 2009. Blue carbon: the role of healthy oceans in binding carbon: a rapid response assessment. UNEP/Earthprint.
- Randeniya, L. K., Vohralik, P. F., Plumb, I. C., 2002. Stratospheric ozone depletion at northern mid latitudes in the 21st century: The importance of future concentrations of greenhouse gases nitrous oxide and methane. *Geophysical Research Letters*, 29(4), 10-1.
- Schmidt, A., Hanson, C., Kathilankal, J., Law, B.E., 2011. Classification and assessment of turbulent fluxes above ecosystems in North-America with self-organizing feature map networks. *Agricultural and Forest Meteorology*, 151(4), 508-520.
- Stocker, T. F., Qin, D., Plattner, G. K., Tignor, M., Allen, S. K., Boschung, J., ... Midgley, B. M., 2013. IPCC, 2013: climate change 2013: the physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change.
- Xiao, J., Zhuang, Q., Baldocchi, D. D., Law, B.E., Richardson, A.D., Chen, J.,...Tom, M.S., 2008. Estimation of net ecosystem carbon exchange for the conterminous United States by combining MODIS and AmeriFlux data. *Agricultural and Forest Meteorology*, 148(11), 1827-1847.
- Zhu, Q., Peng, C., Chen, H., Fang, X., Liu, J., Jiang, H., ..., Yang, G., 2015. Estimating global natural wetland methane emissions using process modelling: spatio-temporal patterns and contributions to atmospheric methane fluctuations. *Global Ecology and Biogeography*

Chapter 2: Scaling-based robust empirical modeling of growing season net ecosystem exchange: application for deciduous forests

Abstract

Gap-filling of net ecosystem exchange (NEE) in eddy covariance time series is necessary for the accurate estimation of the terrestrial ecosystems' carbon balance. Available gap-filling techniques are mostly site-specific and require meteorological data (non-linear regression, artificial neural network, semi-parametric models) and historical NEE information (mean diurnal variation method) as inputs. These limitations require the development of a model that is robust in space and time and needs minimum parameter set. Application of scaling can lead to such generalized modeling approach for the gap-filling of different diurnal hours of NEE. In this connection, a scaling based empirical model was developed for spatiotemporally robust predictions of the diurnal cycle of terrestrial growing season NEE from a corresponding single reference observation by using an extended stochastic harmonic algorithm (ESHA). A reference-time observation from each diurnal cycle was utilized as the scaling parameter to normalize the corresponding hourly observations of NEE for different days into a single, dimensionless diurnal curve. The scaled harmonic model was evaluated by predicting the hourly observed growing season NEE time-series of June to September during 2005-2013 for five AmeriFlux sites, representing distinct deciduous forest ecosystems as the case studies. The developed model incorporated a parsimonious set of five parameters, which exhibited spatiotemporal robustness by collapsing into narrow ranges with no apparent trends or groupings among the different days and sites. Model calibrations demonstrated good modeling efficiency (coefficient of determination, $R^2 = 0.68$ to 0.88) and accuracy (ratio of root-mean-square error to the standard deviation of observations, $RSR = 0.35$ to 0.57) for different sites. Model validations with the years of growing season observations outside the calibration period also showed good performance ($R^2 = 0.66$ to

0.79, and RSR = 0.47 to 0.60). Upscaling of the individual forest-specific models to that at the generalized deciduous forest level was performed by averaging the temporal ensemble of model parameters among all study sites; the upscaled model was tested for the same forest sites for the validation years and modeling efficiencies were similar to the calibration and validation results. The empirical NEE model is applicable for simulating continuous (e.g., hourly) NEE time-series from a corresponding single reference observation at the same or similar ecosystem sites. Once the model is parameterized, it needs only single NEE observational data for each diurnal cycle to predict the entire cycle. The method can potentially be used for a robust estimation of missing data in observed time-series of NEE or other periodic ecohydrological variables for diverse terrestrial ecosystems.

2.1 Introduction

Terrestrial net ecosystem exchange (NEE) is a key measure of ecosystem carbon budget (balance between photosynthesis and respiration), representing the overall exchange of CO₂ between the ecosystem and the atmosphere (Law et al., 2002; Xiao et al., 2014). The terrestrial NEE is dictated by daytime photosynthesis (resulting in downward or negative fluxes of CO₂) and nocturnal respiration (upward or positive fluxes), leading to a concave diurnal cycle. However, the NEE cycle can substantially vary from different days, seasons, years and ecosystem sites due to the numerous biological, climatic, hydrological, and ecological processes factors (e.g., ecosystem structure and function, plant type, solar radiation, ambient temperature, soil characteristics, and water availability). The eddy-covariance (EC) method (Baldocchi, 2003, 2008; Aubinet et al., 2012) has led to the emergence of large regional and global network of monitoring towers and databases such as the AmeriFlux and the FluxNET (Baldocchi et al., 2001, Friend et al., 2007), which include measurements of NEE and relevant climate, ecological and environmental variables. The fluxes measured by the EC technique have the temporal resolution

of half-hourly to hourly. However, continuous measurements of the fluxes in the field with such fine resolution are difficult because of the instrument failure, unfavorable weather condition, and a probable breach in the micrometeorological hypothesis. The amount of missing NEE data in AmeriFlux and EuroFlux is 35% in a year as reported by Falge et al. (2001). Proper modeling techniques can estimate these missing measurements (i.e., gap-filling) for an accurate upscaling (instantaneous to weekly, monthly, seasonal time scales) through utilization of the “big ecological data” to derive appropriate management and policy strategies for the enhancement of ecosystem carbon sequestration. A standard method is, therefore, necessary to generate fine-resolution, continuous NEE data from the limited available observations. The site- and day-specific variation in NEE also warrants the development of robust modeling and prediction tools, which can provide a proper reconstruction of different diurnal cycles using the available incomplete information at the same ecosystem sites.

Numerous mechanistic models were developed to predict terrestrial CO₂ exchanges by involving different process drivers and detailed parameterizations of relevant process components (i.e., photosynthesis, respiration, soil biogeochemistry). Examples of process-based models include DLEM (Tian et al., 2010), ED2 (Medvigy, 2009), ORCHIDEE (Krinner et al., 2005), BIOME-BGC (Thornton et al., 2002), HYBRID (White et al., 1999), and ECOSYS (Grant et al., 2012). A comparative evaluation of available biosphere models using data from the North American Carbon Program (NACP) showed overall higher uncertainty in predicting CO₂ exchanges; suggesting a need for improved parameterization at various temporal scales (Schwalm et al., 2010; Deitze et al., 2011; Barr et al., 2013; Stoy et al., 2013; Raczka et al., 2013). Data-driven models have also been developed to fill gaps in measured data of ecosystem CO₂ exchanges; these methods include the mean diurnal variation technique (Falge et al., 2001), environmental conditions based ‘look-up’ tables, (Falge et al., 2001, Reichstein et al., 2005), semi-parametric and non-parametric statistical approaches (Jarvis et al., 2004; Hui et al., 2004;

Stauch and Jarvis, 2006; Gove and Hollinger, 2006; and Menzer et al., 2013), non-linear regressions for photosynthesis and respiration (Michaelis and Menten, 1913; Lloyd and Taylor, 1994; Valentini et al., 1996; and Aubinet et al., 1999), and artificial neural network (Papale and Valentini, 2003; Braswell et al., 2005; Ooba et al., 2006; Schmidt et al., 2008; and Moffat et al., 2010). Papale (2012) summarized the existing gap-filling methods in EC NEE time series and provided a detailed guideline of the NEE time series data treatment. Data-driven predictive models of carbon fluxes were also developed using both remotely sensed and in-situ observations of geographical, environmental and meteorological variables for different ecosystems (e.g., Oechel et al., 2000; Byrne et al., 2005; Wylie et al., 2007; Makela et al., 2008; and Jahan and Gan, 2013). Most of the available data-driven gap-filling techniques (except mean diurnal variation) use meteorological data (e.g., PAR, air temperature, net radiation) as inputs. Even single imputation based look-up table method also utilize different environmental conditions for gap-filling. Only gap-filling method available that is independent of meteorological and environmental data is single imputation based mean diurnal variation (MDV) technique, which relies on historical NEE values to fill gaps.

Several studies have been conducted to compare the performance of the existing gap-filling methods (Falge et al., 2001, Van Wijk et al., 2002, Ooba et al., 2006). Moffat et al. (2007) made a comprehensive comparison among 15 gap-filling methods and reported superior performance of the artificial neural network (ANNs) based techniques. Based on the findings of this study Fluxnet has adopted artificial neural network with pre-sampling and smoothing (ANN_PS), and improved look-up table (marginal distribution sampling) (Reichstein et al., 2005) as standard gap-filling methods (Papale et al., 2012). However, both of the methods are site specific and depend on meteorological and environmental conditions.

Harmonic analysis has been an extensively applied data-driven method to characterize periodic variables (or signals) and develop prediction models by using Fourier series and Fourier transformations. The advantage of harmonic analysis is that it does not require additional environmental variables other than the predictant itself for model development. Several studies have applied harmonic functions for the analysis and modeling of ecosystem CO₂ exchanges. For example, Richardson and Hollinger (2005) and Hollinger et al. (2004) used a second order Fourier regression to predict nocturnal CO₂ exchanges. Griffis and Rouse (2001) applied a power spectral analysis with the fast Fourier transform to analyze and model the inter-annual variability of net ecosystem exchange. Apart from the carbon flux domain, classical harmonic analysis has been utilized to develop empirical models for numerous applications of ecological, environmental and water resources engineering (e.g., Nestler and Long, 1997; Meyers et al., 2001; Dyar and Alhadeff, 2005; and Kumar et al., 2006;).

A notable advantage of data-driven modeling over the process-based counterpart is that the data-driven method can involve a relatively small parameter set and does not necessarily require numerous, complex scientific hypotheses; increasing the scope of obtaining predictions with reduced uncertainty. Further, data-driven models are often much simpler in structure and computationally less expensive. Data-driven models can, therefore, be used as handy tools to estimate ecosystem carbon fluxes and derive appropriate management strategies without requiring expert domain knowledge and heavy computational resources. However, a data-driven model is developed based on certain environmental and metrological conditions; any change of this assumption may require model re-parameterization. A common limitation, however, with the data-driven, as well as the process-based carbon dynamics models is the site-specificity of model parameters, which often requires a new calibration for application at a different site. Moreover, in some cases the interpretability of the data-driven model parameters are more limited than for process-based models.

Scaling is a key concept that can collapse dynamic observations of different references onto a comparable reference platform. Application of scaling in diurnal NEE cycles can lead to the development of a generalized (both in space and time) modeling approach with minimum input requirement for the estimation of missing data. Scaling has been applied in various applications of ecological/environmental science to develop modeling and prediction tools (e.g., West et al., 2001; Milne et al., 2002; O'Connor et al., 2006; Warnaars et al., 2007; and Hondzo and Warnaars, 2008; and). Enquist et al. (2003) applied the scaling concept to build a general model of ecosystem respiration for different terrestrial ecosystems across the globe that provided a platform for the quantitative understanding of the energy and material flux exchanges between the atmosphere and biosphere. Abdul-Aziz et al. (2007a) presented a scaling-based extended stochastic harmonic analysis algorithm (ESHA) to convert stream dissolved oxygen (DO) measured at different clock-times of the day to that at a single referenced time (e.g., noon). Abdul-Aziz and Ishtiaq (2014) utilized the ESHA to develop a scaling based model to predict the entire diurnal cycle of stream DO from a single reference observation. Here the term 'single reference observation' defines a single observational NEE value of the each diurnal cycle that is used for scaling the entire diurnal cycle.

The objective of this paper is to develop a scaling-based robust data-driven model for simulating the diurnal cycle of terrestrial growing season (June to September) NEE from a single reference observation by leveraging the ESHA model of Abdul-Aziz et al. (2007a). It has been hypothesized that scaling of the different diurnal cycles of NEE by the corresponding single reference observations would generate a unique, dimensionless NEE curve for the same site or different sites of similar ecosystems. Fitting of the unique NEE curve with observations is expected to provide an empirical model that can predict high-resolution NEE. Hourly NEE values of different growing season days for different years (2005-2014) from five AmeriFlux deciduous forest sites (Table 2.1, 2.4) are used for the model calibrations and validations. Finally, a

generalized deciduous ecosystem scale model is developed with the ensemble mean of model parameters obtained from all study sites.

2.2 Materials and methods

2.2.1 Scaling concept

Generally NEE should follow a diurnal cycle because of the daily cyclic pattern of photosynthesis and respiration (mainly driven by PAR and temperature). However, during the dormant period (winter and spring) productivity (i.e., photosynthesis) is very insignificant because of the absence of sunlight; this interrupts typical diurnal pattern of NEE at these seasons of the year. On the contrary, growing season (June to September) NEE represents well defined diurnal pattern because of the coupled effect of high daytime productivity and nighttime respiration. The discrepancy among NEE diurnal cycles in different seasons was shown in Figure 2.1a. The plot of hourly NEE of Morgan Forest for 61st (representing dormant period) and 158th (representing growing period) Julian day provided a clear difference in NEE cycles among seasons. Figure 2.1b depicted the time-averaged growing season (June to September) NEE having well defined diurnal cycles (oscillating around zero) for five different deciduous forests (Table 2.1).

The scaling based harmonic modeling method is applicable for those days that have a clear diurnal pattern. To demonstrate the scaling-based modeling concept, a schematic of hypothetical diurnal NEE cycles representing different days (1, 2, ..., N) (N= number of days) for a single or multiple ecosystem sites is shown (Figure 2.2a). Ideally a classical harmonic model (Priestley, 1981) would require a unique set of parameter values to represent each day-specific NEE cycle, leading to N parameter sets for N diurnal cycles. Following ESHA (Abdul-Aziz et al., 2007a), each diurnal cycle is normalized (i.e., scaled) by a corresponding reference-time (t_{ref}) single observation (NEE_{ref}); the scaling should ideally collapse different NEE cycles onto a

general, dimensionless diurnal NEE curve (NEE^*) with a value of unity (i.e., $NEE^*=1.0$) at the t_{ref} (Figure 2.2b). The general NEE curve can be parameterized from the measured NEE data for all the available days with the ESHA estimation framework. The scaled, dimensionless model is inherently robust because it represents all days at the same or different sites with a single set (rather than N sets) of parameters. Finally, the NEE cycles of different days are predicted by multiplying the dimensionless model (NEE^*) by the corresponding single reference observations (NEE_{ref}).

Both the ESHA and a classical harmonic analysis (Priestley, 1981) obtain the best-fit model with all available data; however, the ESHA additionally forces the harmonic model through the reference-time normalized observation (i.e., 1.0). As such, the scaling-based ESHA facilitates the proper representation and reconstruction of the entire diurnal cycle of NEE for different days from the corresponding single reference observations at the same or similar

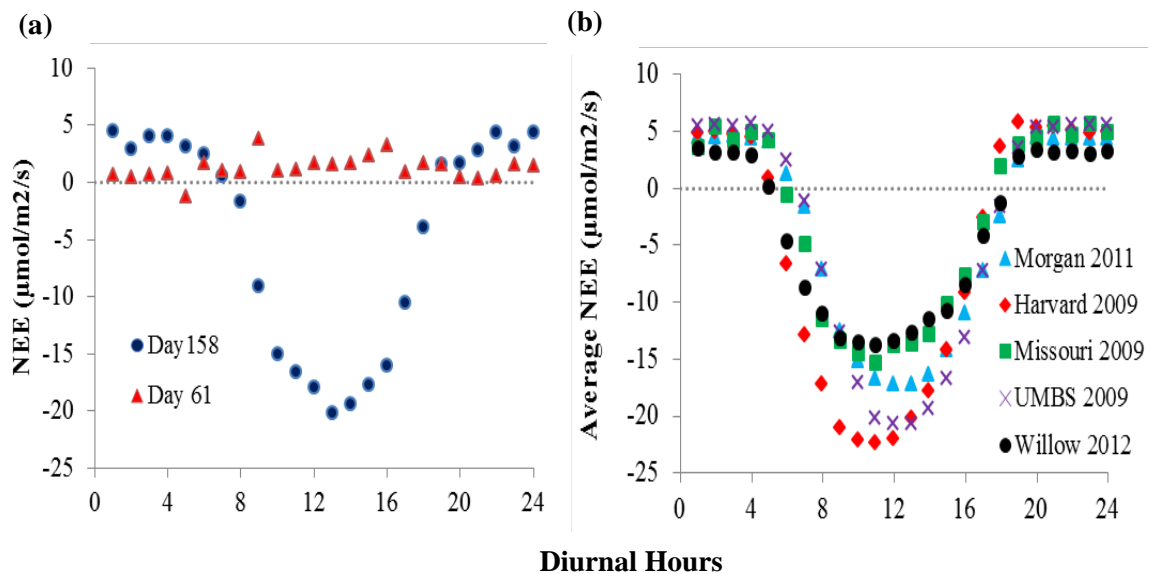


Figure 2.1: Schematics showing (a) comparison of the 24 hour diurnal cycles of NEE for 61th (representing dormant season) and 158th (representing growing season) Julian day of Morgan forest to demonstrate the significant temporal disparity in the shape of the NEE curve and (b) the time averaged growing season (June to September) NEE having well defined diurnal cycles for five different deciduous forests (Table 2.1) considered in the study

Table 2.1: Summary of the selected case study deciduous forests and the associated growing season (June to September) hourly NEE data periods.

Site name	Harvard Forest, MA	Missouri Ozark, MO	Morgan Forest, IN	UMBS Forest, MI	Willow Creek, WI
Location (⁰ N, ⁰ W)	46.08, 89.98	38.74, 92.20	39.32, 86.41	45.56, 84.71	45.81, 90.08
Elevation (m)	340.0	219.4	275.0	234.0	520
Stand age (years)	80.5	77	70	79	66
Canopy height (m)	23	24.2	27	21	24.3
Mean annual temperature (degree C)	6.62	12.11	10.85	5.83	4.02
Mean annual precipitation (mm)	1071	986	1032	803	787
Climate (Koeppen climate indices)	Dfb (Warm Summer Continental)	Cfa (Humid Subtropical: mild with no dry season, hot summer)	Cfa (Humid Subtropical: mild with no dry season, hot summer)	Dfb (Warm Summer Continental)	Dfb (Warm Summer Continental)
Dominant Species	<i>Quercus rubra</i> , <i>Acer rubrum</i>	<i>Quercus alba</i>	<i>Acer saccharum</i> , <i>liriodendron tulipifera</i> , <i>Sassafras albidium</i> , <i>Quercus alba</i> , <i>Quercus nigra</i>	<i>Populus grandidentata</i> , <i>Populus tremuloides</i>	<i>Acer saccharum</i> , <i>Tilia Americana</i> , <i>Fraxinus pennsylvanica</i>
Terrain type	Hilly	Flat to rolling uplands	Ridge/ravine	Gentle slope, glacial drift	Rolling to undulating oval-shaped ridges
Soil type	Glacial till, acidic, low fertility, high organic content	Weller silt loam, acidic to neutral	mesic typic Dystrochrepts	Entic Haplorthod, sandy	Coarse glacial till
Reference	Urbanski et al., 2007; Dang et al., 2011	Gu et al., 2006, 2007	Schmid et al., 2000; Dragoni et al., 2011	Schmid et al., 2003	Cook et al., 2004; Desai et al., 2005

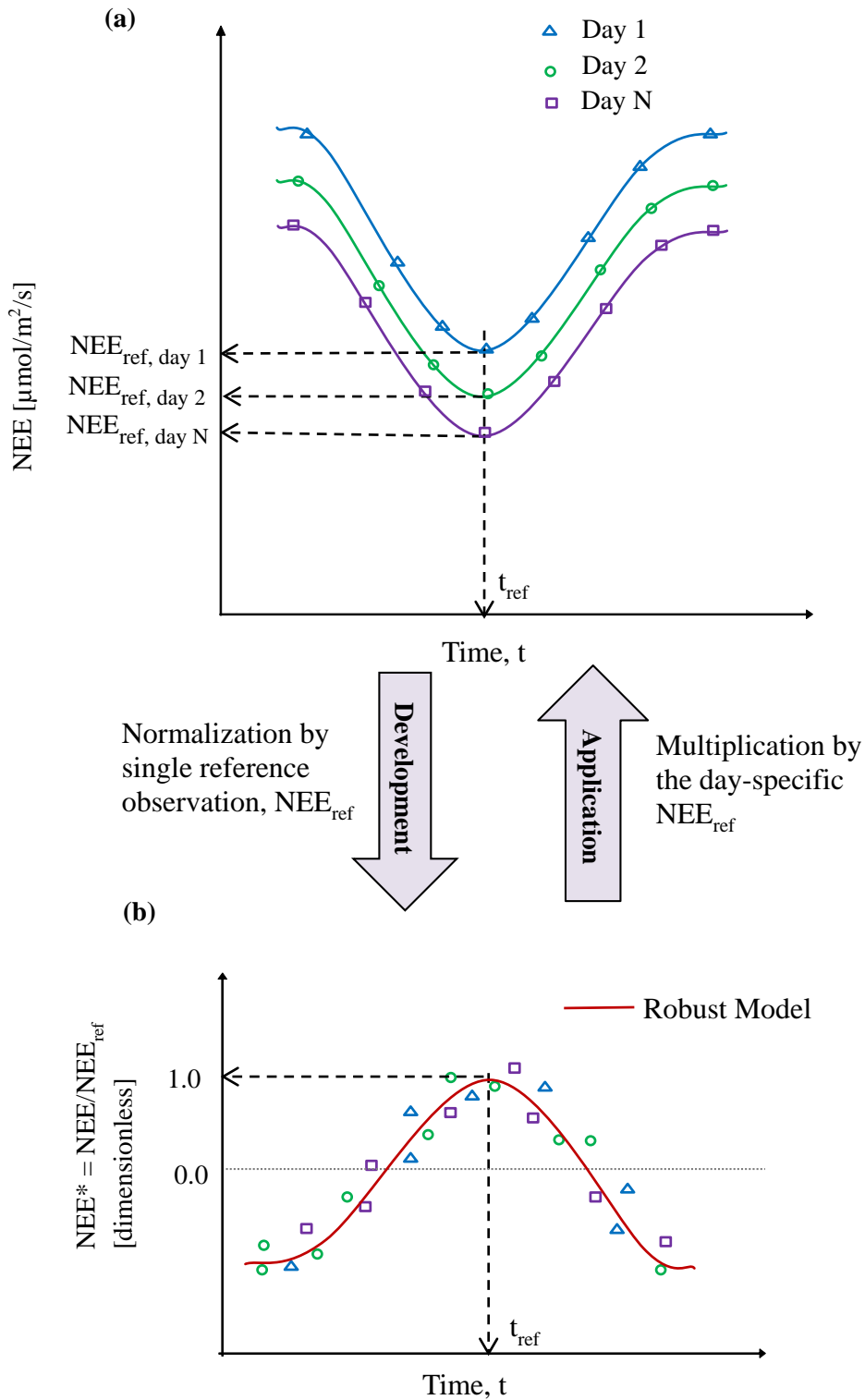


Figure 2.2: Schematic of the conceptual framework for development and application of the scaled net ecosystem exchange (NEE) model.

ecosystem sites. This is useful because (1) NEE data of fine temporal resolutions (e.g., hourly, half-hourly) often have gaps due to instrumental errors or unfavorable weather conditions (Falge et al., 2001), and (2) NEE for many ecosystems in the U.S. and around the world are not yet continuously measured and monitored; posing a challenge to the accurate estimations of carbon budget at the longer time-scales (e.g., weekly, monthly, seasonally). Once the model is parameterized with data from the same or similar ecosystems (e.g., deciduous forests), the scaling-based ESHA model can be utilized to estimate the entire diurnal cycles of NEE for other days and sites by using the corresponding limited observations. Summary of the parameter estimation framework along with the system of solving equations are presented in the Appendix.

2.2.2 Steps of model development and application

The sole input required to develop an ESHA model is a continuous (e.g., hourly) NEE data set representing the full diurnal cycle (24 hours) for at least a single day (and preferably for multiple days) at a single site or multiple sites. The model development steps can be stated as follows:

1. Obtain observed NEE data, $NEE(t)$ for 24-hour cycles at an hourly time (t) intervals.
2. Choose a reference time (t_{ref}) for each cycle to identify the corresponding reference observation, $NEE_{ref}=NEE(t_{ref})$.
3. Scale each NEE cycle by normalizing $NEE(t)$ with the corresponding day-specific reference value to derive a dimensionless cycle as $NEE^*(t)=NEE(t)/NEE_{ref}$.
4. Apply ESHA to fit the $NEE^*(t)$ cycles by forcing the harmonic process through the reference-time normalized observation of 1.0; i.e., $h(t_{ref})=NEE^*(t_{ref})=k=1$.
5. Obtain a generalized scaled model $NEE^*_{mod}(t)$ with the ensemble estimates of daily parameters.

The parameterized dimensionless model of NEE_{mod}^* , representing all days, are then multiplied by the corresponding single reference observations, NEE_{ref} to predict the individual diurnal cycles (NEE_{mod}) for the same site or different sites as follows:

$$NEE_{mod}(t) = NEE_{ref} \times NEE_{mod}^*(t) \quad (1)$$

2.2.3 Selection of optimal number of harmonics

The Selection of the optimal number of harmonics (W) is an important aspect of ESHA to obtain a critical balance of model parsimony, simplicity, and prediction accuracy. The optimal number of harmonics was selected from the calibration dataset based on an Akaike Information Criterion (AIC), which is a well-established measure of the relative performance of a data-driven model in terms of accuracy (Akaike, 1974). Minimum value of the AIC for a particular harmonic value indicates the optimum number of harmonics. Following Abdul-Aziz et al. (2007a), a modified AIC (normalized by the total number of observations) was defined to bring data from different sites on a comparable scale as

$$AIC_{NEE^*}(W) = \ln\left(\frac{SSE_{NEE^*}}{N}\right) + \frac{2}{N}(2W + 1) \quad (2)$$

where W = total number of harmonics, $SSE_{NEE^*} = \sum_{i=1}^N (NEE_i^* - NEE_{mod,i}^*)^2$ = total sum of squared deviations between the normalized observations (NEE_i^*) and the corresponding predictions ($NEE_{mod,i}^*$) assuming the mid-point of diurnal cycle as the reference time (i.e., $t_{ref} = 12$ hours), and N = number of observational data considering all diurnal cycles.

2.2.4 Measurement of model sensitivity and uncertainty

Dimensionless sensitivity coefficients for different model parameters were analytically derived to determine the responses of predicted NEE for any changes in parameters for any day and site.

The sensitivity measures $S_{\hat{a}_0}$, $S_{\hat{a}_k}$, and $S_{\hat{b}_k}$, respectively, for \hat{a}_0 , \hat{a}_k , and \hat{b}_k are stated as

follows:

$$S_{\hat{a}_0}(t) = \frac{\partial NEE_{\text{mod}}(t)}{\partial \hat{a}_0} \bigg/ NEE_{\text{ref}} = 1.0 \quad (3)$$

$$S_{\hat{a}_k}(t) = \frac{\partial NEE_{\text{mod}}(t)}{\partial \hat{a}_k} \bigg/ NEE_{\text{ref}} = \cos(2\pi f_k t) - \cos(2\pi f_k t_{\text{ref}}) \quad (4)$$

$$S_{\hat{b}_k}(t) = \frac{\partial NEE_{\text{mod}}(t)}{\partial \hat{b}_k} \bigg/ NEE_{\text{ref}} = \sin(2\pi f_k t) - \sin(2\pi f_k t_{\text{ref}}) \quad (5)$$

where $k = 1, 2, \dots, W$ indicates the parameter sensitivity coefficients for the oscillatory model components (e.g., $S_{\hat{a}_1}$, $S_{\hat{b}_1}$, $S_{\hat{a}_2}$, and $S_{\hat{b}_2}$ for a two harmonic model) and t refers to any hour of the diurnal cycle. Normalization of the sensitivity parameters by the day- and site-specific single reference observations (NEE_{ref}) ensured their comparability for any day and site. Furthermore, a combined first order uncertainty analysis was conducted (Eq. 6) by applying Taylor series expansion (Mays, 2011) and keeping only the first order terms; model uncertainty was measured by estimating the standard deviation of the predicted NEE with a simultaneous change of parameters by their respective standard deviations as follows:

$$\sigma_{NEE_{\text{mod}}}(t) = NEE_{\text{ref}} \left[\sigma_{\hat{a}_0}^2 S_{\hat{a}_0}^2 + \sum_{k=1}^W \sigma_{\hat{a}_k}^2 S_{\hat{a}_k}^2(t) + \sum_{k=1}^W \sigma_{\hat{b}_k}^2 S_{\hat{b}_k}^2(t) \right]^{1/2} \quad (6)$$

where W is the optimal number of harmonics (e.g., $W = 2$ for a two harmonic model); $\sigma_{\hat{a}_0}$, $\sigma_{\hat{a}_k}$, and $\sigma_{\hat{b}_k}$ = standard deviations of the model parameters; and t = any hour of the diurnal cycle of any day and site.

2.2.5 Case study sites and data sets

Performance evaluation of the model requires a careful selection of study sites that represent similarities as well as distinctive diversities. Based on the availability of continuous (e.g., hourly) eddy covariance measurements, hourly NEE data from June to September of different years for five broad-leaf deciduous forest sites of the AmeriFlux network (AmeriFlux website, 2014) were selected (Table 2.1, Figure 2.3) for the modeling. The case study sites that ranged from the upper mid-western to the northeastern U.S.A. were the (1) Harvard Forest, Massachusetts, (2) Missouri Ozark, Missouri, (3) Morgan Monroe State Forest, Indiana, (4) UMBS Forest, Michigan, and (5) Willow Creek, Wisconsin. Although the selected sites represent a similar vegetation cover, they incorporated a relatively moderate variability in carbon flux exchanges and land-atmospheric feedback through the topography, forest stand age, climate, soil type, and hydrology, and ecology (Table 2.1). Selected sites have two types of climate with a moderate gradient in elevation, terrain, forest stand age, mean annual temperature, and precipitation.

The data sets represent both hourly “gap-filled” and “with-gap” NEE ($\mu\text{mol}/\text{m}^2/\text{s}$) of different diurnal (24-hour, starting 1.00 AM) cycles from June to September of different years (2005-2014) collected from AmriFlux (<http://ameriflux.ornl.gov/>) database (Table 2.2).

According to the AmeriFlux sign conventions, positive sign represents upward fluxes (land/forest to atmosphere) of NEE and negative sign denotes downward fluxes. “Gap-filled” NEE data were used for the site-specific parameter estimation and calibration because entire diurnal cycles of different days are required to simulate the ESHA framework. However, once the parameter set is

estimated, the model no longer requires complete diurnal cycles for prediction; only different, day specific single values of NEE (NEE_{ref}) are needed to simulate the NEE cycles. With the estimated mean parameter set for each forest, site- specific model validation for two separate growing seasons (i.e., prediction) was performed with the “with-gap” data that ensures the applicability of the model. Moreover, “with gap” data of the validation years were used to test the performance of an upscaled generalized deciduous ecosystem model (mean of the estimated parameters of all the sites); this additional layer of analysis established the relevance of this model in gap-filling of NEE cycles even though model parameters were determined from the “gap-filled” data.

Two-step data filtering procedures were applied to ensure the quality of the NEE data set. Firstly, extreme values and data outliers were removed based on the interquartile range (IQR) criteria (Turkey, 1977). Any hourly NEE observations outside the range obtained from $Q_1 - 1.5 * IQR$ and $Q_3 + 1.5 * IQR$ (here, Q_1 = first quartile, Q_3 = third quartile, and $IQR = Q_3 - Q_1$) were removed from the dataset for each forest site. Gaps created from such removals were filled through the linear interpolation. Secondly, since one of the critical assumptions of the model is the requirement of well-defined diurnal cycles of NEE, outlying diurnal cycles (i.e., cycles that are not well-defined) were identified and removed based on the daily estimated parameter values that were outside their corresponding 3rd and 97th percentiles (refer to section 2.3.1 for detail description) to avoid bias toward any outlying diurnal cycles for each site. Additionally, to remove the nighttime and early morning venting effect from the NEE data for the southeast (SE) wind direction at the Willow Creek site (Cook et al., 2005), growing season NEE time series were additionally filtered for the SE wind (90-180° from true north) and gaps were filled through linear interpolation.

Different four years of hourly filtered data were used for model calibrations and parameter estimation (Harvard Forest: 2007-10, 415 diurnal cycles; Missouri Ozark: 2009-12,

426 diurnal cycles; Morgan Forest: 2009-12, 411 diurnal cycles; UMBS: 2008-11, 410 diurnal cycles; Willow Creek: 2005-06, 2011-12, 432 diurnal cycles). Similarly, data for different (independent of calibration years) two growing seasons were used for site-specific and ecosystem scale generalized model validations; e.g., Harvard Forest (2011, 2012), Missouri Ozark (2008, 2013), Morgan Forest (2008, 2013), UMBS (2006, 2012), and Willow Creek (2013, 2014). Further, data for the photosynthetic active radiation (PAR), air temperature (TA), and soil water content (SWC) were obtained from the AmeriFlux network for the calibration periods at different study sites (Table 2.2) to explore any relationships of the model parameters with the important drivers of NEE. The data for hourly NEE and associated climate/environmental variables would

incorporate the intra-, as well as inter-annual variability of the growing season carbon fluxes and their major drivers in evaluating the model performance.

Figure 2.3: Study area and geographic locations of the selected AmeriFlux deciduous forest sites.

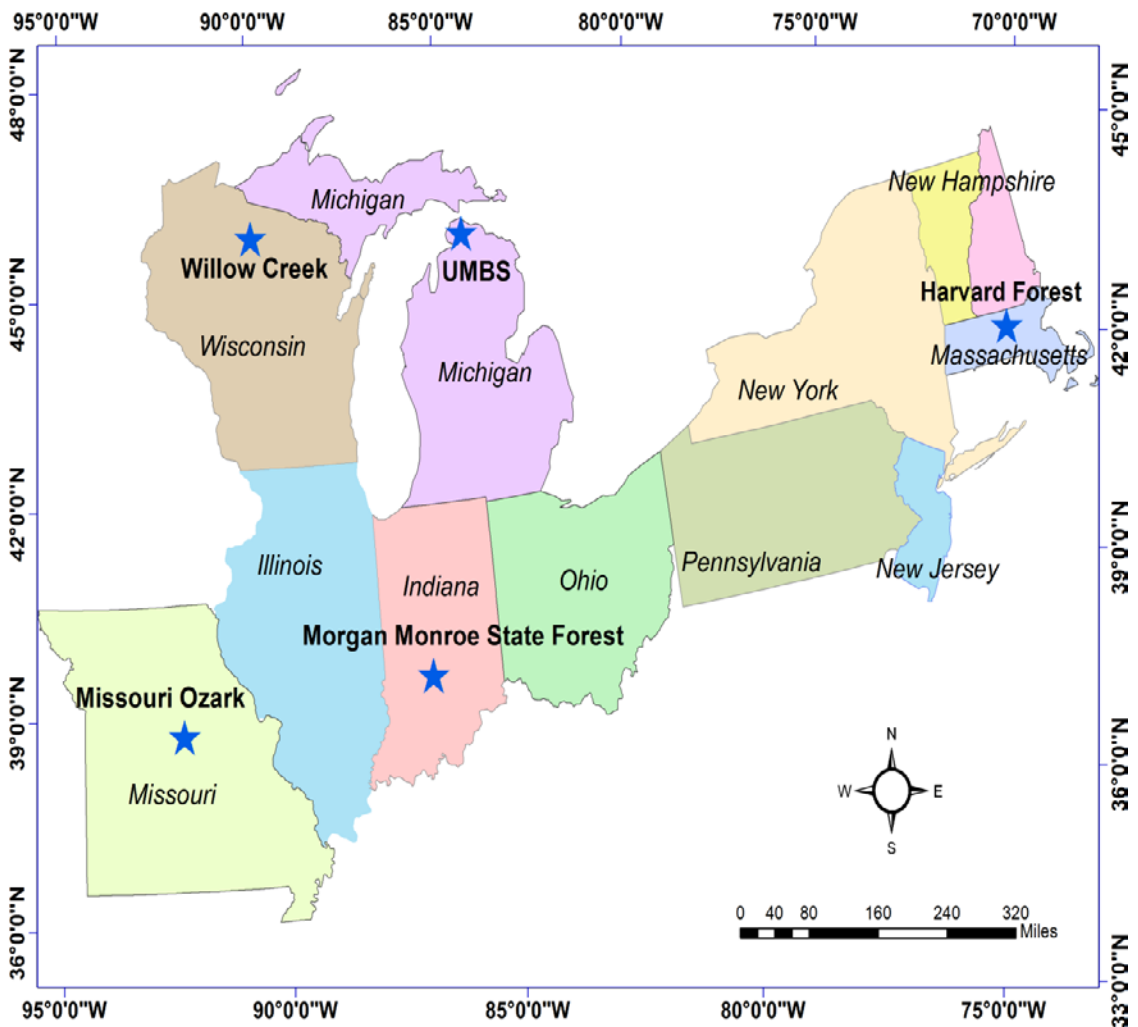


Table 2.2: Growing season (June to September) averages of NEE and associated environmental/meteorological variables for the calibration years for the study sites

Site Name	Calibration years	NEE ($\mu\text{mol}/\text{m}^2/\text{s}$)	PAR ($\mu\text{mol}/\text{m}^2/\text{s}$)	Air temperature ($^{\circ}\text{C}$)	Soil Water Content (%)
Harvard	2007-10	-5.34	303.44	7.90	-
Missouri	2008-11	-3.61	657.38	23.40	35.16
Morgan	2009-12	-3.65	528.08	23.11	26.26
UMBS	2008-11	-3.46	439.84	18.04	7.01
Willow Creek	2005-06, 2011-12	-4.59	485.32	18.19	28.77

2.3 Results and discussion

2.3.1 Optimal number of harmonics

Minimum AIC_{NEE^*} (Equation 5) resulted at $W = 2$ for nearly all the sites (Figure 2.4). Only for the Missouri Ozark Forest, AIC_{NEE^*} for three harmonics (i.e., $W = 3$) was slightly (0.05%) smaller than that of two harmonics; however, the slight accuracy increased the number of model parameters by 40% (5 and 7 parameters for, respectively, 2 and 3 harmonics). To keep a desired balance of model parsimony and accuracy, $W = 2$ was selected as the optimal number of harmonics for all the study forests, leading to the estimation of five parameters (\hat{a}_0 , \hat{a}_1 , \hat{a}_2 , \hat{b}_1 , and \hat{b}_2) at each site. Abdul-Aziz et al. (2007a) also considered an optimal number of two harmonics for their stream DO scaling model using the same ESHA method.

2.3.2 Optimal reference time

The theoretical formulation of ESHA (Abdul-Aziz et al., 2007a) allows anytime from the 24-hour NEE cycle to be the reference-time (t_{ref}) to identify the corresponding single observation as the scaling operator. Practically, however, it is important to assess the model response to the different choices of t_{ref} and identify the desirable set, if any, for an optimal model performance. Utilizing the observed data for different study sites with $W = 2$, coefficient of determination, R^2 was used to measure the impacts of different reference-times on the quality of model predictions. As shown

(Figure 2.5), all the sites moderate to high sensitivity to the choice of t_{ref} . Overall, the maximum prediction efficiency fell into the window of 10-15 hours (i.e., 10:00 AM to 3:00 PM local time) for different study sites. The results also suggest that scaling different diurnal cycles by the corresponding midday window (10-15 hours) reference observations can provide an adequately high modeling efficiency for all stations.

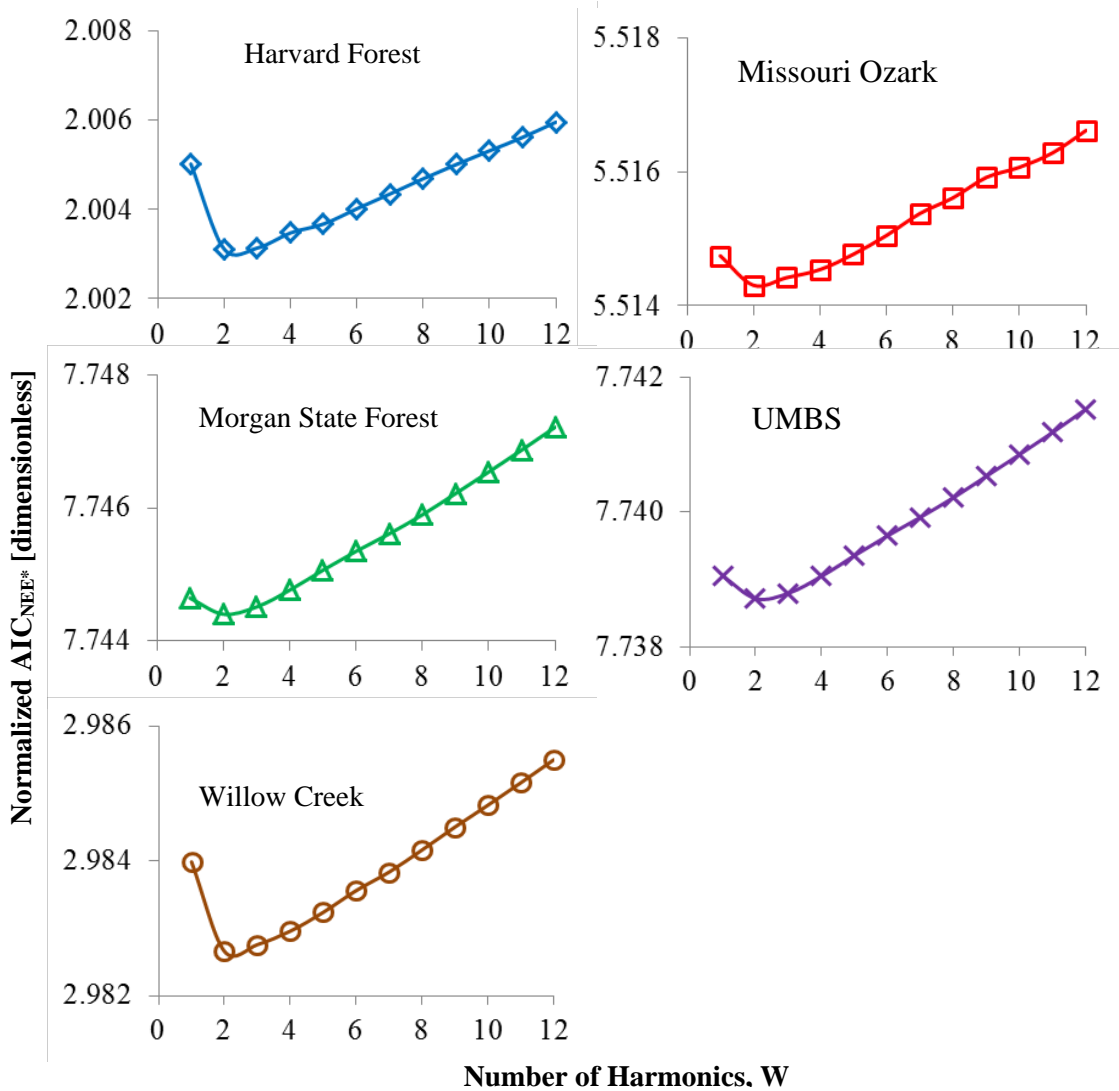


Figure 2.4: Plots of the normalized AIC_{NEE*} with the number of harmonics (W) for different study forests.

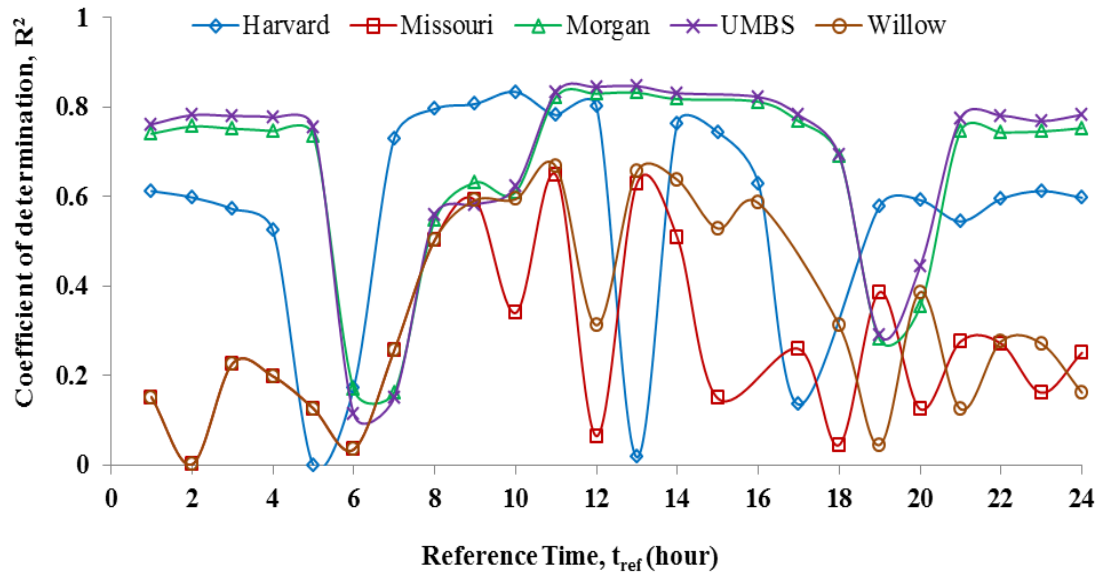


Figure 2.5: Impact of choosing different reference time (t_{ref}) on the modeling efficiency, R^2 for the five study forests.

The finding is similar to that of Abdul-Aziz et al. (2007a) who used midday as the t_{ref} for parameterizing the diurnal cycles of DO in different streams. Based on the calculated R^2 , different t_{ref} were selected for different stations within the optimal model fitting range (i.e., 10:00 AM to 1:00 PM local time) to demonstrate the flexibility of selecting reference-time with ESHA. As such, 10 hours (i.e., 10:00 AM) for Harvard forest, 11 hours (i.e., 11:00 AM) for Missouri, 13 hours (i.e., 1:00 PM) for Morgan, UMBS, and Willow Creek were selected as the reference-times (Tables 2.4, 2.5).

2.3.3 Model evaluations for individual deciduous forests

Estimation of model parameters in time and space

Five dimensionless parameters (\hat{a}_0 , \hat{a}_1 , \hat{a}_2 , \hat{b}_1 , and \hat{b}_2) of the scaled NEE model (NEE^*_{mod}) were estimated for the calibration days (June-September for four years) for each of the five deciduous forest sites. To omit the outliers, site-specific daily parameter values outside their

corresponding 3rd and 97th percentiles were removed from further analysis; this filtering led to the exclusion of around 12-15% of the daily parameter values and corresponding diurnal NEE cycles for each site. The temporal ensemble (June-September) set of parameters were obtained by averaging the filtered daily estimates over the filtered calibration days (Table 2.3) at each site; the spatiotemporal ensemble parameters (i.e., a generalized set to represent all the study sites) were calculated by averaging the corresponding June-September means of the five forests over the calibration period.

Table 2.3: Temporal ensembles (multi-year site-specific averages over June-September) and spatiotemporal means of daily estimated model parameters for the five deciduous forests.

Site name	Parameterization year	\hat{a}_0	\hat{a}_1	\hat{a}_2	\hat{b}_1	\hat{b}_2
Harvard Forest, MA	2007-10 (415 days)	0.235 (0.075)	-0.660 (0.094)	0.134 (0.076)	0.149 (0.080)	-0.060 (0.080)
Missouri Ozark, MO	2008-11 (426 days)	0.196 (0.062)	-0.669 (0.101)	0.145 (0.096)	-0.057 (0.095)	-0.035 (0.115)
Morgan Forest, IN	2009-12 (411 days)	0.191 (0.059)	-0.668 (0.082)	0.163 (0.068)	-0.062 (0.071)	0.015 (0.086)
UMBS Forest, MI	2008-11 (420 days)	0.174 (0.085)	-0.690 (0.102)	0.158 (0.087)	-0.062 (0.080)	0.014 (0.087)
Willow Creek, WI	2005-06, 2011-12 (432 days)	0.272 (0.093)	-0.678 (0.106)	0.102 (0.114)	0.013 (0.130)	-0.024 (0.104)
Spatiotemporal Mean		0.214	-0.673	0.140	0.019	-0.018

Notes: Notes: (1) Values in parentheses are the standard deviations of the estimated parameters; (2) Hat (^) sign represents fitted or estimated parameters of the harmonic (Fourier) series.

Among the different sites, the constant parameter \hat{a}_0 (the non-oscillating, major model component representing a quasi-average daily NEE*) varied from 0.177 to 0.272 with a spatiotemporal mean of 0.213. The first harmonic parameters of \hat{a}_1 and \hat{b}_1 (representing the primary oscillations of the diurnal cycle) varied, respectively, from -0.655 to -0.692 (with a spatiotemporal mean of -0.670) and from -0.092 to 0.149 (with a spatiotemporal mean of -0.025). The second harmonic parameters of \hat{a}_2 and \hat{b}_2 (representing the secondary oscillations) varied, respectively, from 0.102 to 0.171 (spatiotemporal mean = 0.138) and from -0.060 to 0.025

(spatiotemporal mean = -0.02). The relatively small parameter ranges and standard deviations indicated the spatiotemporal robustness of the dimensionless model parameters and subsequent predictions. Spatiotemporal trends of model parameters were also investigated by plotting the filtered (i.e., within the respective 3rd and 97th percentiles), day-specific estimates of \hat{a}_0 , \hat{a}_1 and \hat{b}_1 with the corresponding parameterization days (June to September for four years) of calibrations for all the study forests (Figure 2.6). In general, the parameters (\hat{a}_0 , \hat{a}_1 and \hat{b}_1) did not show any notable increasing or decreasing trend over the 16 months (four months per year for the four years) within and among the different forest sites. However, parameters showed clustered dispersions for September, 2012 at Missouri site; this is likely caused by the less well-defined diurnal cycles of NEE during transitions with the winter season for this particular time. Overall, the model parameters collapsed into narrow bands among different days and sites, indicating spatiotemporal robustness.

The relationships between the model parameters and the major meteorological/environmental process drivers of NEE were investigated to achieve mechanistic insights into the spatiotemporal robustness of estimated parameters for varying environmental conditions. As an example, the daily estimates of the principal parameter, \hat{a}_0 were plotted with the corresponding 24-hour photosynthetic active radiation (PAR), air temperature (TA) and soil water content (SWC) for different study sites (Figure 2.7; SWC were not available for the Harvard Forest). The daily PAR, TA and SWC from different stations were normalized by their corresponding temporal averages (over June to September for four calibration years) to bring them on a comparable scale. The \hat{a}_0 did not display any noteworthy trend with the PAR, TA and SWC for any forests, reemphasizing the robustness of model parameters. However, the higher dispersion of \hat{a}_0 for the Missouri site suggested more variable response of the quasi-average dimensionless daily NEE to the major process drivers at these stations than that for the other

study sites; this can also explain the higher dispersions of \hat{a}_0 when plotted with the cumulative model days for this site (see Fig. 6). The 1st harmonic parameters (\hat{a}_1 and \hat{b}_1), also did not exhibit any noteworthy trend with the PAR, TA, and SWC (Figure 2. 8 and 2.9) except the Missouri forest, where some clustered dispersion of parameters were observed. The observed dispersion of the parameters at the Missouri site can be attributed to its relatively higher temperature, PAR, and higher soil water content (due the existence of silty loam soils and flat topography) compare to the other sites (Table 2.1, 2.2).

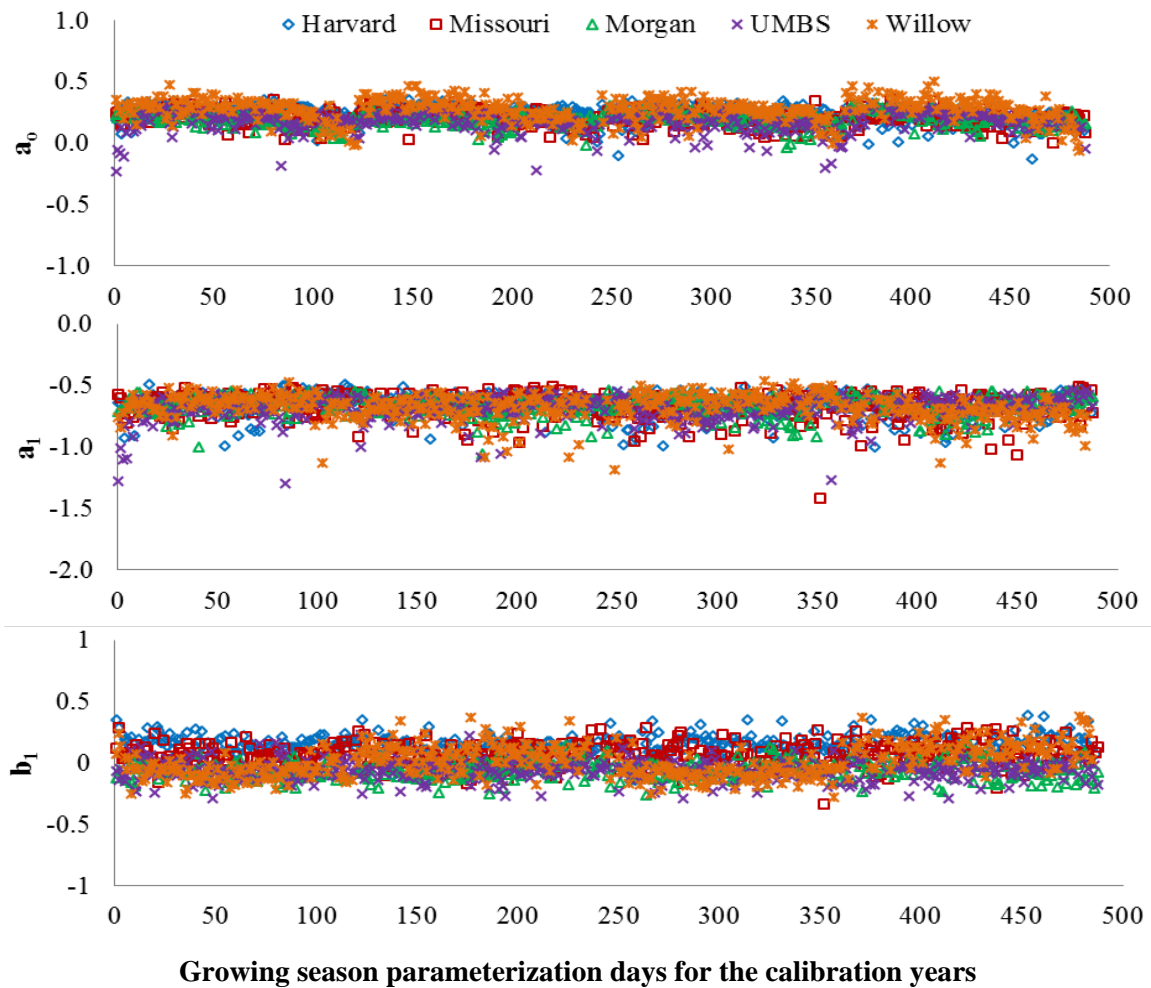


Figure 2.6: Daily time-series of estimated parameters during the four years of calibration for the five deciduous forests, demonstrating the spatiotemporal robustness of the model parameters. The a_0 , a_1 , and b_1 , respectively, refer to \hat{a}_0 , \hat{a}_1 and \hat{b}_1 where the hat ($\hat{}$) represents estimations.

The observed collapse (i.e., robustness) of the estimated parameters is a unique outcome of scaling; normalization of the each diurnal NEE cycle by the corresponding single reference observation brought parameters of different days and sites into the comparable ranges. The impressive spatiotemporal robustness of the model parameters suggests that hourly NEE predictions made by using the June-September aggregations (i.e., site-specific ensemble mean) of daily parameters over the 4-year calibration periods would also be robust in time and space.

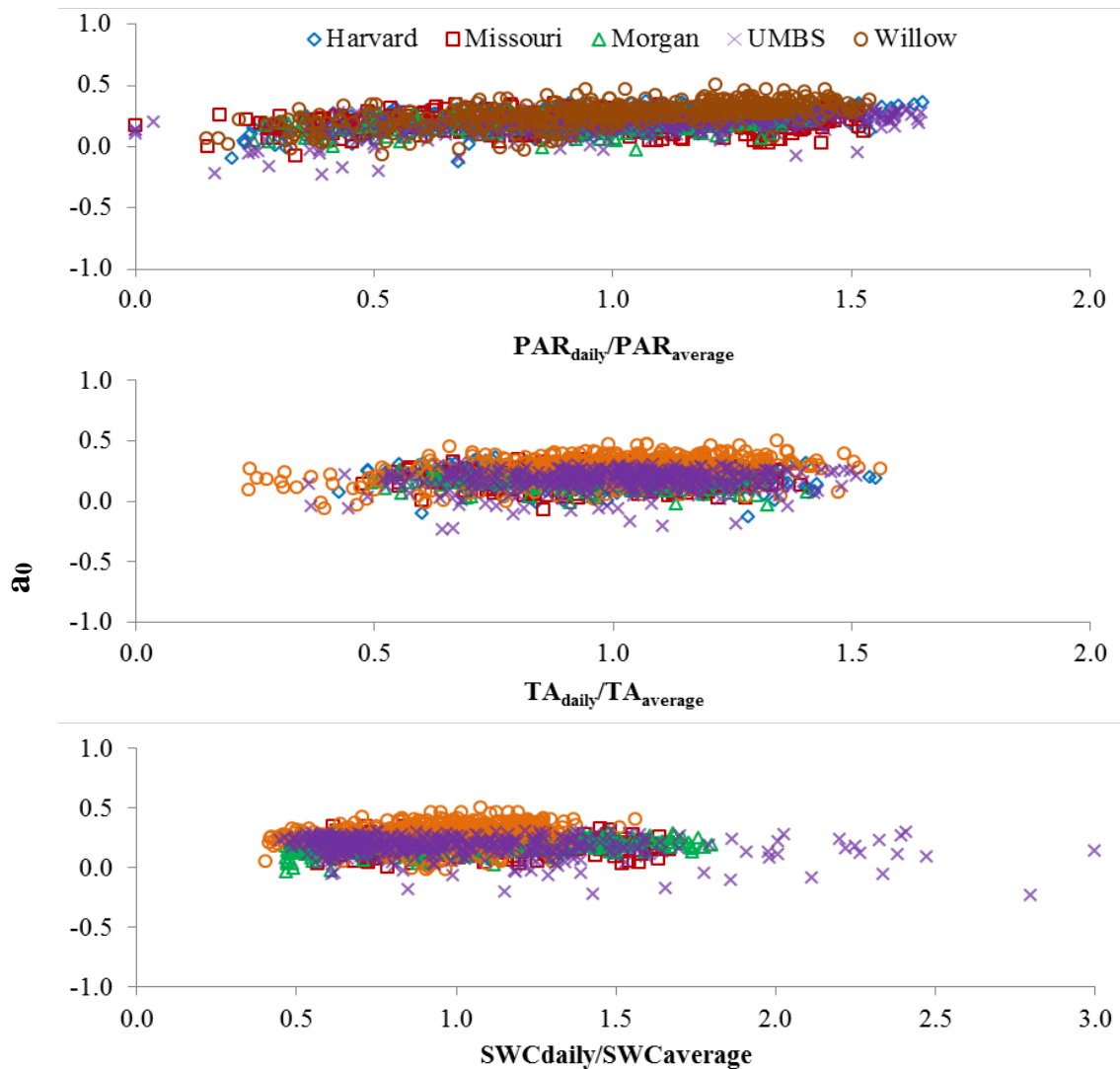


Figure 2.7: Plots of the daily estimates of the principal model parameter, \hat{a}_0 with the corresponding (as normalized by the respective site-specific averages) (a) photosynthetic active radiation ($PAR_{daily}/PAR_{average}$), (b) air temperature ($TA_{daily}/TA_{average}$), and (c) soil water content ($SWC_{daily}/SWC_{average}$) over the four different years of model calibrations for different sites.

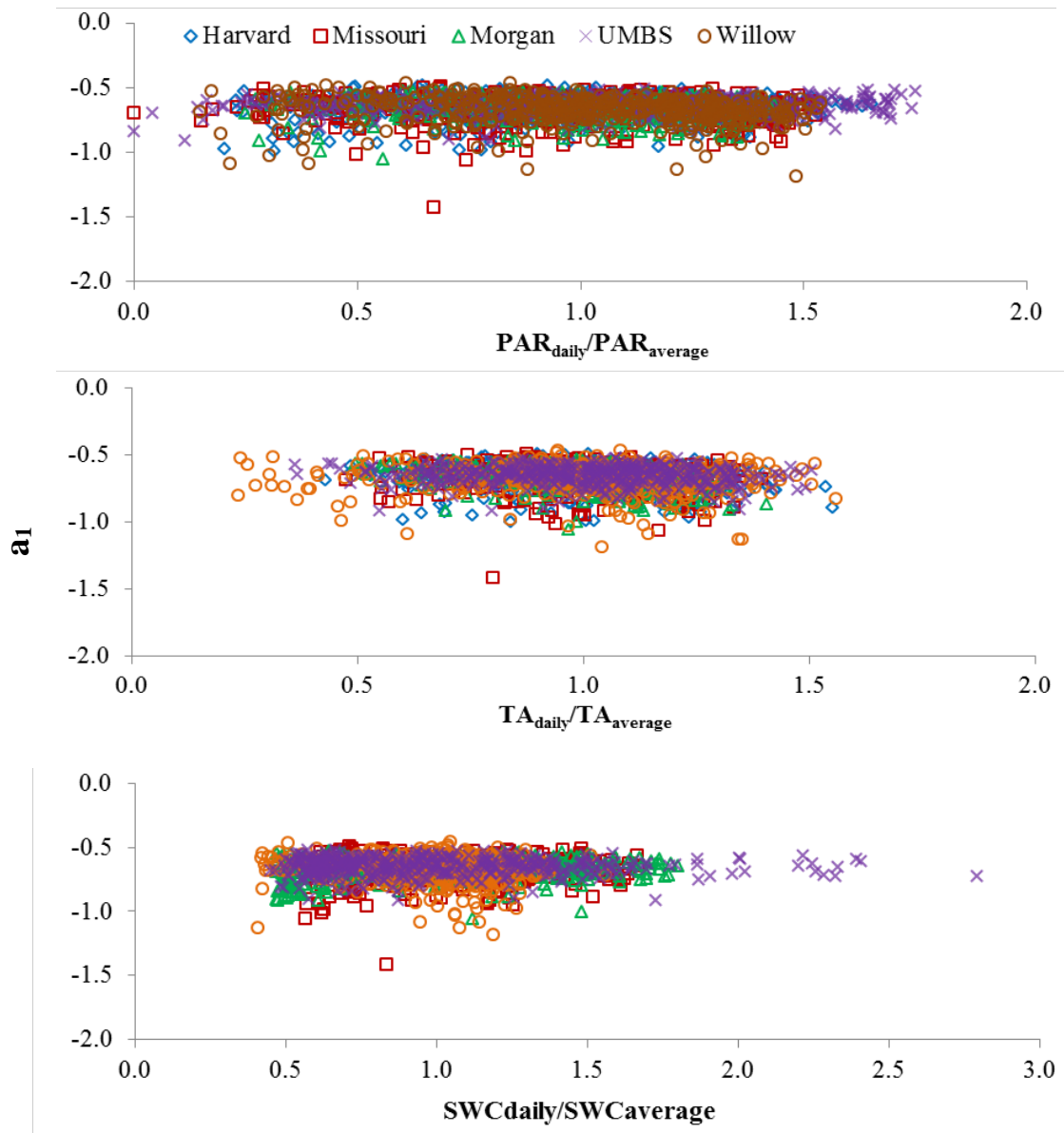


Figure 2.8 Plots of the daily estimates of the principal model parameter, \hat{a}_1 with the corresponding (as normalized by the respective site-specific averages) (a) photosynthetic active radiation ($PAR_{daily}/PAR_{average}$), (b) air temperature ($TA_{daily}/TA_{average}$), and (c) soil water content ($SWC_{daily}/SWC_{average}$) over the four different years of model calibrations for different sites.

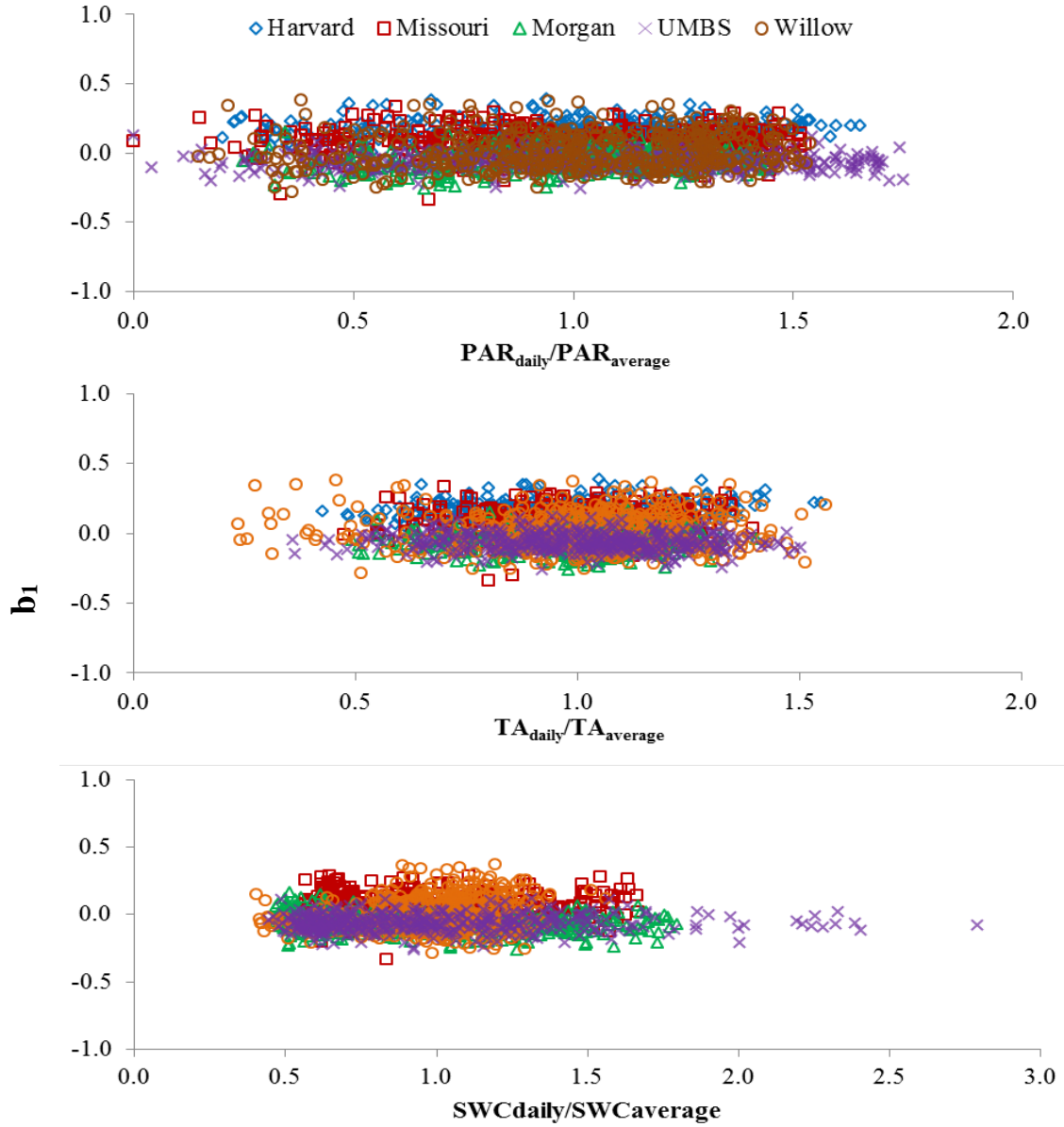


Figure 2.9: Plots of the daily estimates of the principal model parameter, \hat{b}_1 with the corresponding (as normalized by the respective site-specific averages) (a) photosynthetic active radiation ($PAR_{daily}/PAR_{average}$), (b) air temperature ($TA_{daily}/TA_{average}$), and (c) soil water content ($SWC_{daily}/SWC_{average}$) over the four different years of model calibrations for different sites.

2.3.4 Temporally robust predictions from model calibrations and validations

The site-specific, temporally ensemble parameters were used to predict the hourly NEE cycles from the corresponding day-specific single reference observations for the growing season calibration days. Three criteria were used to evaluate the model prediction performance (Table 2.4): (i) ratio of root-mean-square error to the standard deviation of observations (RSR) that indicates model accuracy (Singh et al., 2004); (ii) coefficient of determination (R^2), indicating the amount of observed data variance explained by the model ($R^2=1.0$ indicates a perfect model); and (iii) Nash-Sutcliffe Efficiency (NSE) (Nash and Sutcliffe, 1970) that also measures the goodness of model fit (see Appendix for the mathematical expressions of RSR and NSE). Following the extensive analysis of Moriasi et al. (2007) on model performance, an RSR from 0 to 0.50 indicates a perfect to very good model, from 0.5 to 0.6 indicates a good model, and from 0.6 to 0.7 refers to a satisfactory model; a model with $RSR > 0.70$ is considered unsatisfactory. Similar to the R^2 , the NSE of 1.0 indicates a perfect model, whereas $NSE > 0$ indicates better fitting with the harmonic model than that of the daily mean of hourly observations as an alternative model. Among the five different forests, RSR of the calibrated models varied between 0.35 to 0.557 (good to very good prediction accuracy); R^2 ranged from 0.68 to 0.88 and NSE ranged from 0.68 to 0.88, indicating high model fitting efficiency (Table 2.4).

The calibrated NEE model was validated with independent two years of growing season NEE data among the different study forests (Table 2.4). The NEE diurnal cycles (outside of the respective calibration years) were predicted from their corresponding reference-time single observations by using the previously estimated site-specific, temporal ensemble parameters. The model validation demonstrated nearly equivalent prediction performance ($RSR = 0.47$ to 0.60 ; $R^2 = 0.66$ to 0.79 ; and $NSE = 0.64$ to 0.78), compared to that of the calibrations among the different study sites.

Table 2.4: Results of calibrations and validations for the individual forest-specific NEE models.

		Calibrations				Validations			
Sites	Year	Ref. time (hr)	R ²	RSR	NSE	Year	R ²	RSR	NSE
Harvard Forest	2007-10	10	0.87	0.37	0.86	2011	0.78	0.48	0.77
						2012	0.75	0.50	0.75
Missouri Ozark	2008-11	11	0.71	0.54	0.71	2007	0.64	0.60	0.63
						2013	0.69	0.57	0.68
Morgan Forest	2009-12	13	0.87	0.36	0.87	2008	0.79	0.47	0.78
						2013	0.76	0.50	0.75
UMBS Forest	2008-11	13	0.87	0.36	0.87	2006	0.76	0.51	0.74
						2012	0.76	0.50	0.75
Willow Creek	2005-06, 2011-12	13	0.79	0.47	0.79	2013	0.77	0.48	0.77
						2014	0.75	0.51	0.74

Notes: (1) Data represent June through September each year; (2) Ref. time represents the cumulative hour of a 24-hour clock at which the reference observation was made; (3) RSR is the ratio of root-mean-square error to the standard deviation of observations; and (4) NSE (Nash-Sutcliffe Efficiency) is the measure of the goodness of model fit.

Performance of the model in the dimensional form was evaluated by separating the daytime and nighttime NEE and comparing the predicted means (in $\mu\text{mol}/\text{m}^2/\text{s}$ unit) against observed means for all the sites for both calibration and validation periods (Table 2.6). Daytime and nighttime were defined as periods of the day when photosynthetic active radiation (PAR) was positive (higher than zero), and zero (additionally, PAR corresponding to the nocturnal respiration was set to zero), receptivity (Moffat et al., 2007). The calibration mean differences (i.e., deviation of model mean from the observed mean) between observed and predicted varied from 0.33 to 0.62 $\mu\text{mol}/\text{m}^2/\text{s}$ and 0.01 to 0.55 $\mu\text{mol}/\text{m}^2/\text{s}$ for daytime and nighttime NEE, respectively for all the sites; validation mean deviations were slightly higher than the calibration mean differences (daytime: 0.04 to 1.22 $\mu\text{mol}/\text{m}^2/\text{s}$, nighttime: 0.39 to 0.78 $\mu\text{mol}/\text{m}^2/\text{s}$) (Table 2.6). Overall, the site-specific calibrations and validations illustrated the ability of the scaling-based ESHA model for predicting the entire diurnal cycles of hourly NEE from the respective single reference observations. The success in simulating hourly NEE for June-September during

2005-2014, by using the temporally ensemble (multi-year, multi-season means) single set of site-specific model parameters, also demonstrated the temporal robustness of the scaled model. Since the diurnal NEE cycle varies significantly among different days even at the same site (Richardson et al., 2007), a harmonic model without proper scaling cannot predict the different hourly NEE cycles using a single set of ensemble parameters. Scaling by the day-specific single reference observations resulted in the collapse of different cycles into a general dimensionless diurnal curve (as shown in Figure 2.2) and provided the comparable values of estimated parameters over the entire simulation (calibration and validation) periods; which led to the temporally high quality predictions for each forest. Moreover, since the model uses day-specific NEE observation (i.e., NEE_{ref}) to simulate the entire diurnal cycle, the inter-annual variability of the predicted NEE, if any for a site, would be reflected through NEE_{ref} . The successful validation of the scaled model reemphasized its applicability for predicting the entire diurnal cycles of NEE based on limited available data; aiding to fill gaps in time-series of observed NEE at the study forests.

2.3.5 Spatiotemporal robustness: upscaling to a generalized deciduous forest model

The spatiotemporal robustness of the model predictions (spatiotemporal robustness of the model parameters was demonstrated in section 2.3.3) was tested by upscaling the site-specific hourly NEE_{ref} models to a generalized deciduous forest model; parameters of the upscaled model were obtained by spatially averaging the temporally ensemble, site-specific parameters (previously estimated through calibrations) for all the sites (the last row of Table 2.3). The generalized deciduous model was then used to predict hourly growing season diurnal NEE cycles from their corresponding single reference observations at different study sites for independent (of calibration) years representing two different decades (e.g., the years of 2006 and 2012 for the UMBS; Table 2.5 for further details). The hourly NEE, as predicted by the upscaled generalized model, showed a nearly similar accuracy ($RSR = 0.48$ to 0.58) and model fitting efficiency ($R^2 =$

0.66 to 0.77; NSE = 0.66 to 0.77) compare to the validation results (Table 2.5); comparison of NEE means also resembled the validation and calibration results (daytime mean difference: 0.05 to 1.21 $\mu\text{mol}/\text{m}^2/\text{s}$, nighttime mean difference: 0.03 to 0.51 $\mu\text{mol}/\text{m}^2/\text{s}$) (Table 2.6).

For visualization, scattered plots of the observed versus the generalized model predicted NEE for different years between 2006 and 2013 (independent of the calibration years) for the five study forests were presented (Figure 2.10). Similar to the performance of the site-specific models (calibrations and validations), the quality of predictions from the generalized model was notably good for all the sites except the Missouri Ozark. The scaled harmonic model could not entirely predict the extreme peaks (carbon emissions) and troughs (sequestrations) of NEE, resulting in a reduction of modeling performance. The possible reasons could be (i) the existence of higher parameter sensitivity and modeling uncertainty in the nighttime hours away from the reference hour (Figure 2.11, 2.12) that is being reflected in the predicted extreme values, (ii) use of the temporally and spatiotemporally averaged parameters (day-specific model parameters would make far better predictions for extremes), (iii) random sampling errors, which are a function of averaging time (Baldocchi, 2003) and statistical error from gap filling (Falge et al., 2001), and (iv) inability of the model to predict instantaneous perturbations amid the moderate geographical gradients in climate, topography, soil type, hydrology, forest stand age and ecology among the five deciduous forests (Table 2.1). The predictions of different diurnal NEE cycles for all the study sites using a single parameter set demonstrated a high spatiotemporal robustness of the dimensionless, scaled model. The overall results indicate that the scaling-based generalized ESHA model can be utilized to predict diurnal NEE cycles at other deciduous forest ecosystems (which are not included in this study) if the corresponding day-specific single reference observations are available. However, incorporation of data from more deciduous forests could further improve the robustness of the generalized model by enhancing the ecosystem-based spatial coverage and domain of the scaled NEE model.

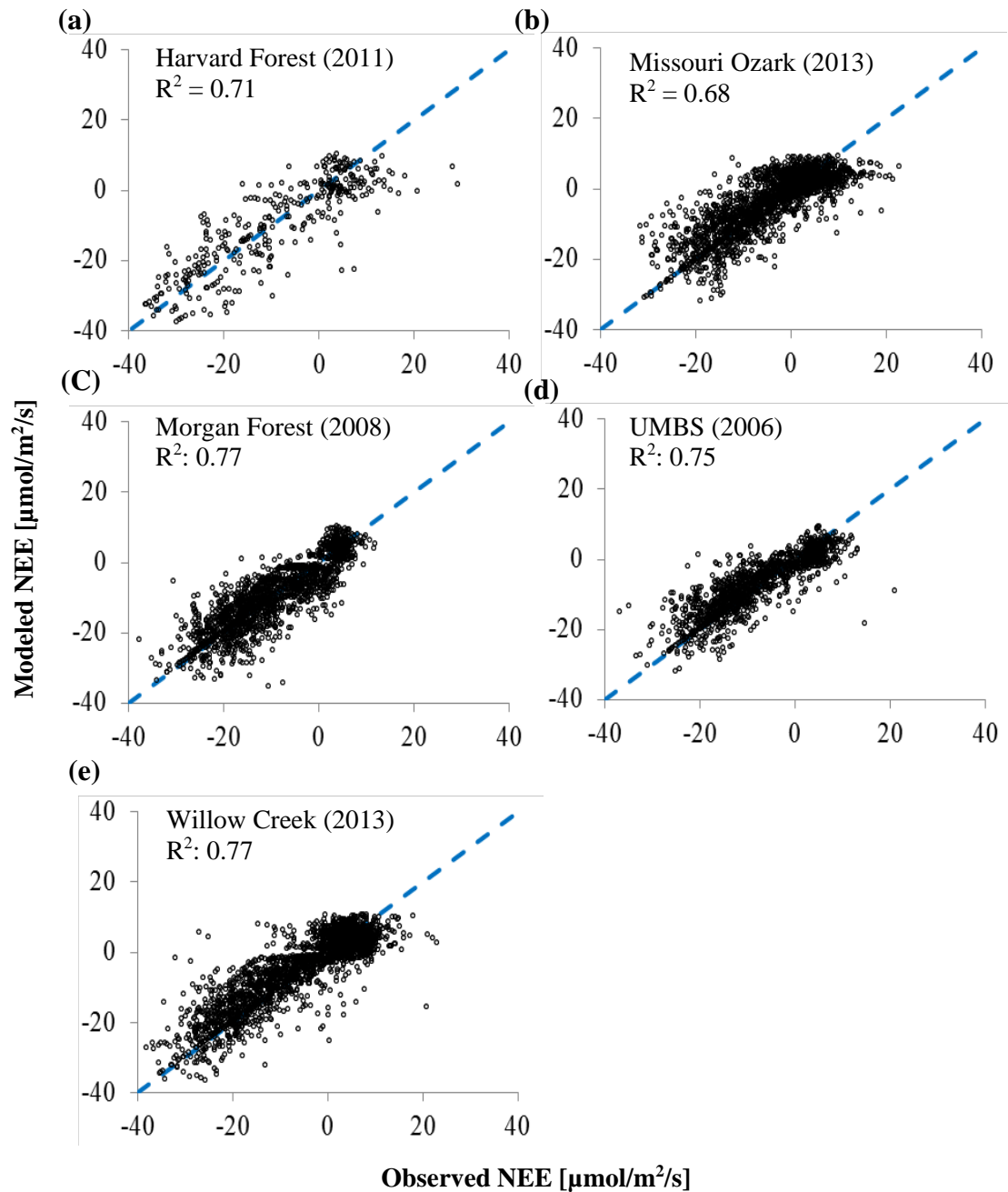


Figure 2.10: Plots of the hourly predicted and observed NEE of different years for the (a) Harvard Forest, (b) Missouri Ozark, (c) Morgan State Forest, (d) UMBS, and (e) Willow Creek Forest using the generalized deciduous forest model. Dashed line represents the 1:1 perfect fit line.

Table 2.5: Performance of the generalized deciduous forest model in predicting the hourly NEE of different days and years for the five study forests.

Sites	Year	Ref. time (hr)	r^2	RSR	NSE
Harvard Forest, MA	2011	10	0.71	0.55	0.70
	2012		0.71	0.54	0.71
Missouri Ozark, MO	2007	11	0.64	0.60	0.64
	2013		0.68	0.58	0.67
Morgan Forest, IN	2008	13	0.77	0.49	0.76
	2013		0.73	0.52	0.73
UMBS Forest, MI	2006	13	0.75	0.51	0.74
	2012		0.75	0.51	0.74
Willow Creek, WI	2013	13	0.77	0.49	0.76
	2014		0.75	0.50	0.75

Notes: (1) Data represent May through October each year; (2) Ref. time represents the cumulative hour of a 24-hour clock at which the reference observation was made; (3) RSR is the ratio of root-mean-square error to the standard deviation of observations; and (4) NSE (Nash-Sutcliffe Efficiency) is the measure of the goodness of model fit.

3.3.6 Comparison of the model with the existing gap-filling techniques

The daytime and nighttime performances of the ESHA model were compared with the root mean square error (RMSE) values of the existing gap-filling techniques reported by Moffat et al. (2007). The calculated daytime RMSE of the ESHA model ranged between 4.10-5.98, 5.52-6.88, and 5.56-7.66 gC/m²/day (1.0 gC/m²/day = 0.96 μmol/m²/s) for the calibration, validation, and upscaled generalized model, respectively for all the sites. Similarly, nighttime RMSE varied from 2.24-6.32, 3.20-5.70, and 3.30-5.87 gC/m²/day for the calibration, validation, and upscaled generalized model, respectively for all the sites. These RMSE values are comparable to the RMSE values of mean diurnal variation (MDV) (Falge et al., 2001), and multiple imputation method (MIM) (Hui et al., 2004) as obtained from Moffat et al. (2007). However, both daytime and nighttime RMSE values of the ESHA model were higher than the sophisticated gap-filling methods like ANNs (artificial neural network), NLRs (non-linear regressions), and MDS (marginal distribution sampling) (Moffat et al., 2007)

Table 2.6: Comparison of the daytime and nighttime predicted means against the observed daytime and nighttime means of NEE for the calibration, validation and generalized model for all the study forests

Site name	Calibration				Validation				Generalized Model			
	Average daytime NEE ($\mu\text{mol}/\text{m}^2/\text{s}$)		Average nighttime NEE ($\mu\text{mol}/\text{m}^2/\text{s}$)		Average daytime NEE ($\mu\text{mol}/\text{m}^2/\text{s}$)		Average nighttime NEE ($\mu\text{mol}/\text{m}^2/\text{s}$)		Average daytime NEE ($\mu\text{mol}/\text{m}^2/\text{s}$)		Average nighttime NEE ($\mu\text{mol}/\text{m}^2/\text{s}$)	
	Obs	Mod	Obs	Mod	Obs	Mod	Obs	Mod	Obs	Mod	Obs	Mod
	Harvard Forest	9.77	9.44	4.53	5.02	12.18	12.55	5.28	4.60	12.18	12.01	5.28
Missouri Ozark	6.89	6.51	4.62	5.17	4.22	3.32	3.39	3.01	4.22	3.78	3.39	2.89
Morgan Forest	9.84	9.22	4.46	4.47	10.43	9.93	4.05	4.83	10.43	10.38	4.05	4.42
UMBS	7.60	7.10	4.97	5.39	9.19	7.97	4.43	4.82	9.19	8.38	4.43	3.92
Willow Creek	10.48	9.97	3.31	3.52	10.46	10.42	4.33	3.60	10.46	9.25	4.33	4.36

Although the ESHA model did not perform equally with the high ranked gap-filling methods, it has the benefit of predicting hourly NEE values without requiring meteorological inputs and historical NEE values that are necessary for the ANNs, NLRs and look-up tables; rather, the ESHA model requires a single day-specific NEE observation to simulate the entire diurnal cycle of NEE once parameterized.

2.3.7 Model sensitivity and uncertainty

At any diurnal hour, the dimensionless sensitivity coefficient for \hat{a}_0 (the non-oscillatory, principal model component parameter) was 1.0; referring to a constant, linear rate of change in predicted NEE for any change of \hat{a}_0 . The sensitivity coefficients for the oscillatory model component parameters (\hat{a}_k and \hat{b}_k) varied dynamically with the diurnal hours due to the presence of sine-cosine functions (Figure 2.11; see Equation 4-5). Among the different diurnal hours at the five study forests, the sensitivity coefficients of the first harmonic parameters (\hat{a}_1 and \hat{b}_1) varied, respectively, between -0.13 to 1.96 and -1.5 to 1.25, whereas the second harmonic parameters (\hat{a}_2 and \hat{b}_2) varied between -1.87 to 0.5 and -0.5 to 1.86. The oscillatory component parameters did not show any sensitivity at the reference time since the harmonic process (the dimensionless diurnal curve) was forced to pass through the reference-time observation.

Model uncertainty due to the individual parameter uncertainty was computed by separately changing each parameter value by the corresponding standard deviation (Table 2.3). The Morgan forest was considered as an example to demonstrate the computed model uncertainty; since the reference-time observation (NEE_{ref}) in Equations (3-5) can be different on different days, I used average $NEE_{ref} = -17.27 \mu\text{mol}/\text{m}^2/\text{s}$ for simplicity and convenience of

demonstrating model responses to the parameter changes for any day. For the Morgan forest, changing the temporal mean of \hat{a}_0 (0.191) by the corresponding standard deviation (0.059) resulted in a change of $-1.02 \mu\text{mol}/\text{m}^2/\text{s}$ in the predicted NEE; changing the harmonic parameters ($\hat{a}_1, \hat{b}_1, \hat{a}_2, \hat{b}_2$) by their respective standard deviations resulted in absolute changes of the predicted NEE by up to $-2.78, -1.54, -2.18,$ and $-2.23 \mu\text{mol}/\text{m}^2/\text{s}$, respectively. However, an overall measure of the model uncertainty for all study forests was obtained by computing the standard deviations of the predicted NEE for simultaneously changing all random parameters by their corresponding standard deviations. Based on the first-order uncertainty measure (Equation 6), for the site-specific average (from on the growing season data used for calibration) of the reference observations (average $NEE_{ref} = -16.56$ to $-21.11 \mu\text{mol}/\text{m}^2/\text{s}$), standard deviations of the predicted NEE for different diurnal hours varied between -1.06 and $5.07 \mu\text{mol}/\text{m}^2/\text{s}$ among the different forests (Figure 2.12). The 95% confidence limits of the uncertainties were also shown in Figure 2.12.

The analytical measure of model sensitivity and uncertainty is an essential attribute of the scaled NEE model. The results showed that the prediction uncertainties of the scaled harmonic model were higher in case of the simultaneous perturbations in parameters than the standalone perturbations. Moreover, uncertainties were high for all the sites during the nighttime hours as time deviates from the reference time. Since, the diurnal cycles are predicted from the day-specific noontime (i.e. 10:00 AM to 1:00 PM) reference observations (Equation 1), prediction uncertainties propagate from the reference time, which also explains the inability of the model in predicting extreme NEE fluxes as discussed in section 2.3.5. Although for simplicity and ease of demonstration, the average values of reference NEE for each site was used for Equation (4-6), the NEE_{ref} can be different for different days and sites.

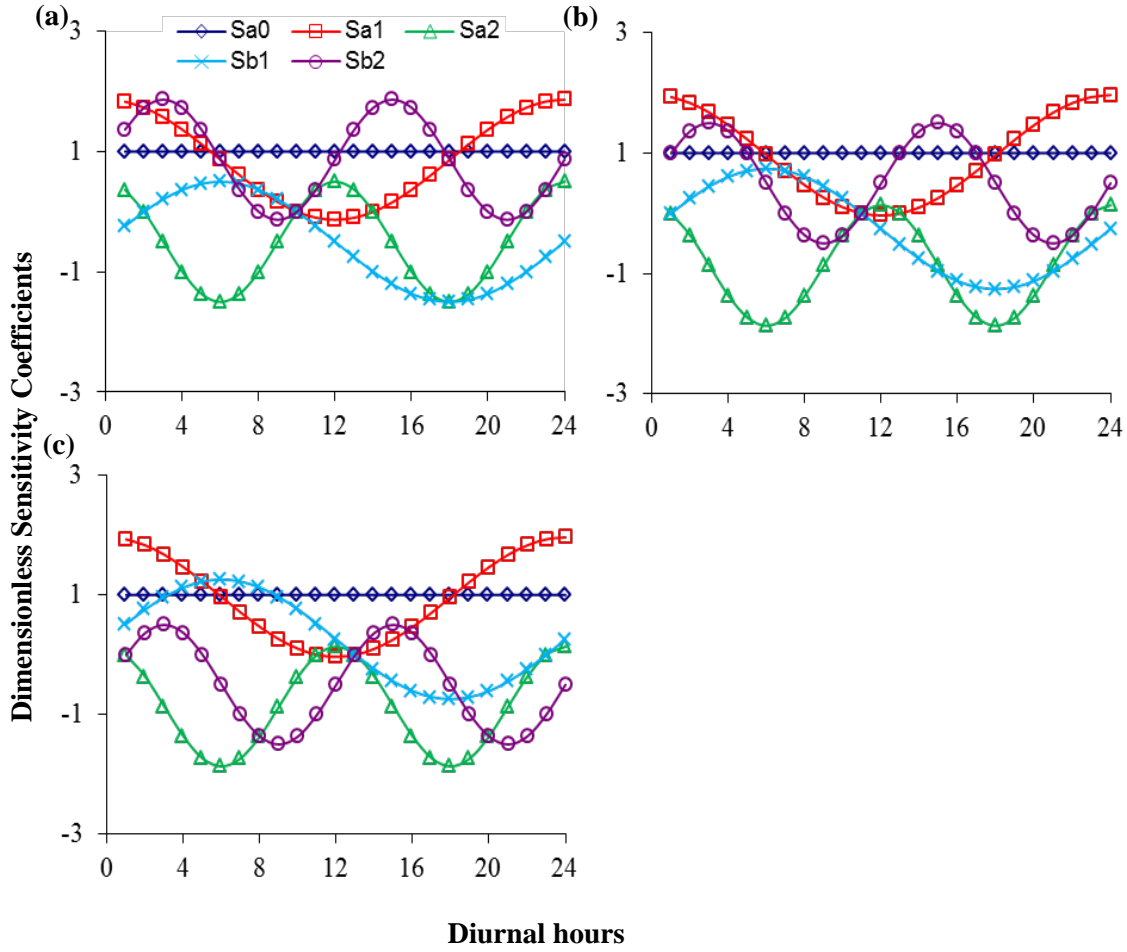


Figure 2.11: Plot of dimensionless parameter sensitivity coefficients ($S_{\hat{a}_0}$, $S_{\hat{a}_1}$, $S_{\hat{b}_1}$, $S_{\hat{a}_2}$, and $S_{\hat{b}_2}$) with the hours of any diurnal cycle, indicating the model sensitivity for the (a) Harvard Forest, (b) Missouri Ozark and Willow Creek, (c) Morgan State Forest, and UMBS Forest (Sa0, Sa1, Sa2, Sb1, and Sb2, respectively, refer to model sensitivity to \hat{a}_0 , \hat{a}_1 , \hat{a}_2 , \hat{b}_1 and \hat{b}_2).

2.3.8 Gap-filling of NEE time-series and predictions in relatively ungauged forests

Success in constructing the diurnal NEE cycles from a single reference observation indicated the potential of the scaling-based harmonic NEE model for filling gaps in high resolution (e.g., half-hourly, hourly) time-series of growing season NEE, as well as for predicting the entire diurnal

cycles of growing season NEE from available limited observations for relatively ungauged forests of similar ecosystem type. The flux towers in historically monitored forests can encounter around 35% missing data annually at the half-hourly scale (Falge et al., 2001), which poses a challenge to achieve accurate estimations of carbon budget at the longer time-scales (e.g., weekly, monthly, seasonally). The available gap-filling methods such as the empirical mean diurnal variation (MDV) and the semi-empirical non-linear regressions and artificial neural networks (see Falge et al., 2001 and Moffat et al., 2007 for details) are generally site-specific in applications (i.e., values or parameters derived for one site cannot be transferred to a different site). Further, the MDV method can fill gaps at a particular time if the corresponding NEE data for 7 to 14 neighboring days are available (Falge et al., 2001); the semi-empirical regressions and ANNs can be used only if the corresponding data for the relevant mechanistic predictors (e.g., radiation, temperature) are available. None of the existing methods can leverage the limited available data for one time instant to estimate the missing NEE for other times within the same or different diurnal periods. In contrast, the scaling-based harmonic NEE model can provide spatiotemporally robust estimations of missing data and predictions of the entire diurnal cycle of NEE based on a corresponding single NEE measurement made at the reference time (mostly noon; can be flexibly chosen between 10:00 AM and 3:00 PM). Once the model is parameterized with continuous NEE data (of preferably multiple days) at the same or similar ecosystems (e.g., different deciduous forests), the scaled NEE model can be utilized to estimate the entire diurnal cycles of NEE for other days and sites by utilizing the corresponding limited observations.

2.3.9 Model generalization and applicability in different ecosystems

Every forest ecosystem has some innate climatic, physiological, biogeochemical and ecological characteristics; the differing characteristics pose a formidable challenge to the generalization of

prediction models among different forests. In contrast, the estimated parameters of the scaling-based harmonic NEE model showed notable robustness in time (June-September of four different years) and space (five different deciduous forests) (Figure 2. 6). Moreover, absence of any mean dependence between the crucial drivers (PAR, temperature, and SWC) of ecosystem carbon exchanges and NEE) reiterated the robustness of the parameters (Figure 2.7,2.8, and 2.9). The spatiotemporal robustness of model parameters ultimately led to the successful spatial

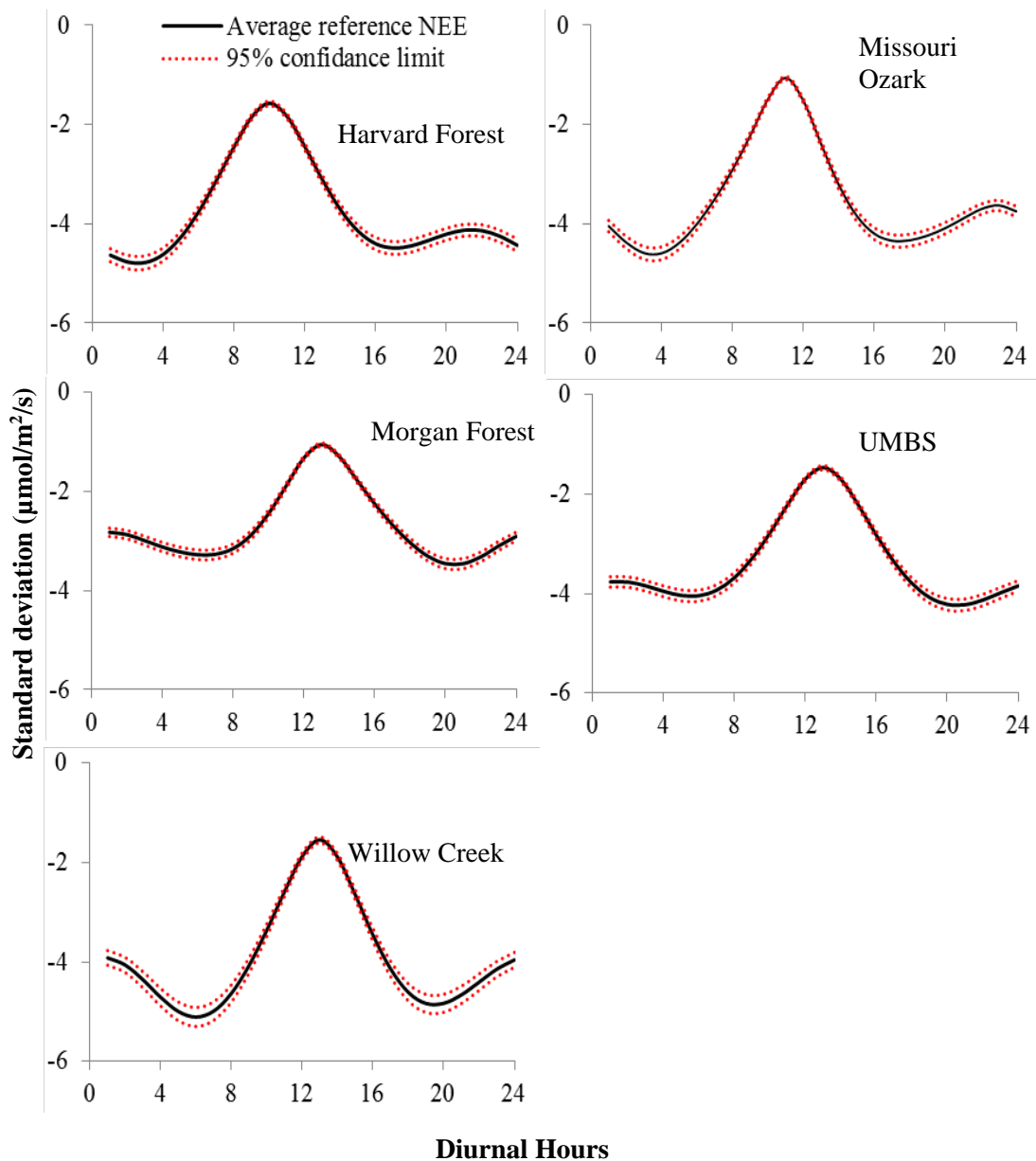


Figure 2.12 Standard deviations, estimated from eqn. 6, of the predicted hourly NEE as an overall measure of the model uncertainty for simultaneously changing all parameters by their respective standard deviations at the five deciduous forests.

aggregation and generalization of the scaled NEE model among the five different deciduous forests. Future studies should investigate whether the scaling-based harmonic (ESHA) model can provide acceptable predictions of the diurnal NEE cycles for other types of ecosystems. If parameters estimated for diverse ecosystem types collapse into narrow ranges, as in the case of different deciduous forests, a scaling-based generalized NEE model for many ecosystems may not be impossible to achieve.

2.3.10 Data reliability and model limitations

The ESHA model parameterization requires at least one (and preferably multiple) 24-hour diurnal cycle(s) of NEE for the calibration period. I used level-2 gap-filled NEE data of AmeriFlux sites to estimate the parameters of the empirical NEE model. These gaps were filled by the AmeriFlux data community mostly by using the artificial neural network (ANN) and MDV methods (see <http://ameriflux.ornl.gov/>), which could introduce input data bias and uncertainty in the ESHA model evaluations. However, the evaluation data set included gap-filled data at different clock-hours (mostly during the night time) for different days, while the day-time NEE contributed the most part of the diurnal variation of NEE. Impact of data gap-filling bias was also theoretically reduced by incorporating a common (two) harmonic model structure for different days during the model evaluation period (2005-2013) among different stations. Any input gap-filled data bias on parameter estimation was further reduced practically (in an inverse manner) by averaging the daily estimated parameters temporally (May-October) and spatiotemporally (among the five study forests) during the four years of model calibrations (i.e., parameter estimation periods) in order to obtain, respectively, the site-specific and the generalized prediction models for the deciduous forests. The calibrated ensemble models can be utilized to obtain spatiotemporally robust predictions and gap-filling of the high resolution (e.g., hourly) NEE diurnal cycles (particularly those not included in the calibration) from their corresponding single reference observations.

Reliability of the AmeriFlux NEE data is another crucial aspect for evaluation of the ESHA model. Schmidt et al. (2012) reported 8.2% relative error (random sampling errors for specific averaging times) for carbon dioxide fluxes of the AmeriFlux measurements; other studies (Berger et al., 2001; Soegaard et al., 2000) reported around 7% or less error during the day and 12% at night. Presence of outliers, particularly at the reference time, would lead to inaccurate scaling of the diurnal NEE cycles, ultimately providing unreliable predictions of the hourly NEE (Grove and Hollinger, 2006). The extreme outliers were removed from the NEE time series based on IQR criteria. Further, to minimize the impact of outliers on predictions, the estimated daily parameters of the harmonic model were filtered by retaining values within their respective 3rd and 97th percentile ranges. Although the filtering excluded 12-15% of daily parameters from the calculation of the ensemble parameters for each forest, the ensemble parameters were utilized to predict all diurnal cycles of the model evaluation periods (including the calibration days filtered by the 3rd to 97th parameter percentiles criteria). The impressive prediction performance of the site-specific and the generalized deciduous forest models, in hindsight, indicate the effectiveness of the applied parameter filter to reduce the impact of outliers on the overall model predictions.

The temporally and spatiotemporally ensemble harmonic models could not accurately predict the extreme values of NEE for some study forests (Figure 2.10); this can be seen as a limitation of temporal and spatiotemporal aggregations (averaging) of the harmonic model parameters. However, the day-specific parameters of the harmonic model can more accurately predict the extreme values of NEE. Furthermore, reliable prediction of high extremes is also unlikely from a conventional process-based model (see the “Introduction” for examples of process models). The extreme values could be better predicted using statistical models specifically designed to predict extremes.

2.4 Conclusions

A scaling-based empirical model was developed by utilizing an extended stochastic harmonic analysis algorithm (ESHA) to obtain spatiotemporally robust predictions of the diurnal cycles of growing season terrestrial NEE from the corresponding single reference observations. Model evaluations with data from five deciduous forests during 2005-13 showed a parsimonious set of five dimensionless parameters, which exhibited spatiotemporal robustness by collapsing into relatively narrow ranges (i.e., no clear trends or groupings) among different days and sites. The model parameters were also insensitive to the individual forest-specific crucial drivers (PAR, temperature, and SWC) of ecosystem carbon exchange. Impressive predictions of the hourly NEE for different days and years with an ensemble seasonal (June-September of four calibration years) set of model parameters demonstrated the temporal robustness of the scaled harmonic model. The analytically derived parameter sensitivity coefficients and uncertainty measure showed a notable robustness of the model predictions to the standalone and simultaneous perturbations in parameters. An upscaled, generalized deciduous forest model showed similar performance to that of the individual forest-specific models; demonstrating the spatiotemporal robustness of the scaled NEE model in predicting the diurnal NEE cycles for different forest sites.

The relatively robust empirical NEE model, once parameterized, can be applied for simulating continuous (e.g., hourly) NEE time-series from a single reference observation (or a set of limited observations) at the same or similar ecosystem sites without requiring day- or site-specific calibrations. The method can potentially be used for a robust estimation of missing data and gap-filling in observed time-series of NEE or other periodic ecohydrological variables (e.g., latent heat flux, sensitive heat flux) for diverse terrestrial ecosystems. Since the harmonic model can construct the entire diurnal cycle of NEE, it can lead to an accurate estimate of the longer temporal-scale (e.g., weekly, monthly, seasonally) ecosystem carbon budget; aiding to derive

appropriate strategies for green carbon management. Currently, the model is limited to the growing season NEE because of the requirement of well-defined diurnal cycles; however future work should investigate the possibility of the inclusion of dormant season in the model. A possible option could be the use of day-specific multiple reference observations to force the model to predict the daily NEE cycles, which are not well-defined.

The scaling-based ESHA algorithm (Abdul-Aziz et al., 2007a) to predict the diurnal NEE cycles is a novel approach in ecological modeling. The spatiotemporal robustness of model parameters, as well as the good predictions from the individual and the generalized deciduous forest models, highlighted the effectiveness of the scaling method to unravel and utilize the underlying ecological similitude (i.e., similarities leading to the parametric reduction) for achieving a useful generalization among different days and forests, which represented a moderate geographical gradient in climate, topography and land uses, soil type, hydrology, forest stand age and ecology. The ecological similitude and scaling merits further research in order to achieve robust estimations and predictions of the carbon budget for diverse ecosystems across a range of spatial and temporal scales.

Acknowledgement

The research was funded by grants from the National Science Foundation CBET Environmental Sustainability (Award No. 1336911), and the National Oceanic and Atmospheric Administration through NERRA (Grant No. NA09NOS4190153). The supports are thankfully acknowledged.

References

- Abdul-Aziz, O.I., Wilson, B.N., Gulliver, J.S., 2007a. An extended stochastic harmonic analysis (ESHA) algorithm: Application for dissolved oxygen. *Water Resources Research*, 43, doi:10.1029/2006WR005530.
- Abdul-Aziz, O.I., Wilson, B.N., Gulliver, J.S., 2007b. Calibration and validation of an empirical dissolved oxygen model. *Journal of Environmental Engineering*, 133(7), 698-710.
- Abdul-Aziz, O. I., Ishtiaq, K. S. 2014. Robust empirical modeling of dissolved oxygen in small rivers and streams: Scaling by a single reference observation. *Journal of Hydrology*, 511, 648-657.
- Akaike, H. 1974. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6), 716-723.
- AmeriFlux site and data exploration system website, 2014 <http://ameriflux.ornl.gov/>
- Aubinet, M., Grelle, A., Ibrom, A., et al. 1999. Estimates of the annual net carbon and water exchange of forests: the EUROFLUX methodology. *Advancement in Ecological Research*, 30, 113-175.
- Aubinet, M., Vesala, T., Papale, D., 2012. *Eddy covariance: a practical guide to measurement and data analysis*. Springer Science & Business Media.
- Baldocchi, D. D., 2008. "Breathing" of the terrestrial biosphere: Lessons learned from a global network of carbon dioxide flux measurement systems. *Australian Journal of Botany*, 56(1), 1-26.
- Baldocchi, D. D., 2003. Assessing the eddy covariance technique for evaluating carbon dioxide exchange rates of ecosystems: past, present and future. *Global Change Biology*, 9(4), 479-492.
- Baldocchi, D. D., Falge, E., Gu, L., et al., 2001. FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. *Bulletin of American Meteorological Society*, 82(11), 2415-2434.
- Barr, A. G., Ricciuto, D. M., Schaefer, K., Richardson, A., Agarwal, D., Thornton, P. E., ... Verma, S., 2013. NACP Site: Tower Meteorology, Flux Observations with Uncertainty, and Ancillary Data. Data set. Available on-line [<http://daac.ornl.gov>] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA <http://dx.doi.org/10.3334/ORNLDAAC/1178>.
- Braswell, B.H., Sacks, W.J., Linder, E., Schimel, D.S., 2005. Estimating diurnal to annual ecosystem parameters by synthesis of a carbon flux model with eddy covariance net ecosystem exchange observations. *Global Change Biology*, 11(2), 335-355.
- Byrne, K.A., Kiely, G., Leahy, P., 2005. CO₂ fluxes in adjacent new and permanent temperate grasslands. *Agricultural and Forest Meteorology*, 135(1), 82-92.

- Cook, B. D., Davis, K. J., Wang, W., Desai, A., Berger, B. W., Teclaw, R. M., ... Heilman, W., 2004. Carbon exchange and venting anomalies in an upland deciduous forest in northern Wisconsin, USA. *Agricultural and Forest Meteorology*, 126(3), 271-295.
- Dang, X., Lai, C. T., Hollinger, D. Y., Schauer, A. J., Xiao, J., Munger, J. W., ... Ehleringer, J. R., 2011. Combining tower mixing ratio and community model data to estimate regional-scale net ecosystem carbon exchange by boundary layer inversion over four flux towers in the United States. *Journal of Geophysical Research: Biogeosciences* (2005–2012), 116(G3).
- Desai, A. R., Bolstad, P. V., Cook, B. D., Davis, K. J., Carey, E. V., 2005. Comparing net ecosystem exchange of carbon dioxide between an old-growth and mature forest in the upper Midwest, USA. *Agricultural and Forest Meteorology*, 128(1), 33-55.
- Dietze, M. C., Vargas, R., Richardson, A. D., Stoy, P. C., Barr, A. G., Anderson, R. S., ... Weng, E., 2011. Characterizing the performance of ecosystem models across time scales: A spectral analysis of the North American Carbon Program site-level synthesis. *Journal of Geophysical Research: Biogeosciences* (2005–2012), 116(G4).
- Dragoni, D., Schmid, H. P., Wayson, C. A., Potter, H., Grimmond, C. S. B., Randolph, J. C., 2011. Evidence of increased net ecosystem productivity associated with a longer vegetated season in a deciduous forest in south-central Indiana, USA. *Global Change Biology*, 17(2), 886-897.
- Dyar, T. R., Alhadeff, S. J., 2005. Dissolved oxygen characteristics of Georgia streams. *Proceedings, 2005 Georgia Water Resources Conference*. University of Georgia, Athens, U.S.A., 25–27 April.
- Enquist, B. J., Economo, E. P., Huxman, T. E., Allen, A. P., Ignace, D. D., Gillooly, J. F., 2003. Scaling metabolism from organisms to ecosystems. *Nature*, 423(6940), 639-642.
- Falge, E., Baldocchi, D.D., Olson, R., et al., 2001. Gap filling strategies for defensible annual sums of net ecosystem exchange. *Agricultural and Forest Meteorology*, 107(1), 43-69.
- Friend, A. D., Arneeth, A., Kiang, N. Y., et al., 2007. FLUXNET and modelling the global carbon cycle. *Global Change Biology*, 13(3), 610-633.
- Gove, J.H., Hollinger, D.Y., 2006. Application of a dual unscented Kalman filter for simultaneous state and parameter estimation in problems of surface-atmosphere exchange. *Journal of Geophysical Research*, 111, D08S07.
- Grant, R.F., Baldocchi, D.D., Ma, S., 2012. Ecological controls on net ecosystem productivity of a seasonally dry annual grassland under current and future climates: Modelling with ecosys. *Agricultural and Forest Meteorology*, 152, 189-200.
- Griffis, T.J., Rouse, W.R., 2001. Modelling the interannual variability of net ecosystem CO₂ exchange at a subarctic sedge fen. *Global Change Biology* 7(5), 511-530.

- Gove, J. H., & Hollinger, D. Y., (2006). Application of a dual unscented Kalman filter for simultaneous state and parameter estimation in problems of surface-atmosphere exchange. *Journal of Geophysical Research: Atmospheres* (1984–2012), 111(D8).
- Gu, L., Meyers, T., Pallardy, S. G., Hanson, P. J., Yang, B., Heuer, M., ... Wullschleger, S. D., 2006. Direct and indirect effects of atmospheric conditions and soil moisture on surface energy partitioning revealed by a prolonged drought at a temperate forest site. *Journal of Geophysical Research: Atmospheres* (1984–2012), 111(D16).
- Gu, L., Meyers, T., Pallardy, S. G., Hanson, P. J., Yang, B., Heuer, M., ... Wullschleger, S. D., 2007. Influences of biomass heat and biochemical energy storages on the land surface fluxes and radiative temperature. *Journal of Geophysical Research: Atmospheres* (1984–2012), 112(D2).
- Hollinger, D. Y., Aber, J., Dail, B., et al., 2004. Spatial and temporal variability in forest–atmosphere CO₂ exchange. *Global Change Biology*, 10(10), 1689-1706.
- Hondzo, M., Warnaars, T. A., 2008. Coupled effects of small-scale turbulence and phytoplankton biomass in a small stratified lake. *J. Environ. Eng.* 134(12), 954-960.
- Hui, D., Wan, S., Su, B., Katul, G., Monson, R., Luo, Y., 2004. Gap-filling missing data in eddy covariance measurements using multiple imputation (MI) for annual estimations. *Agricultural and Forest Meteorology*, 121(1), 93-111.
- Jahan, N., Gan, T.Y., 2013. Developing a gross primary production model for coniferous forests of northeastern USA from MODIS data. *International Journal of Applied Earth Observations and Geoinformation*, 25, 11-20.
- Jarvis, A. J., Stauch, V. J., Schulz, K., Young, P. C., 2004. The seasonal temperature dependency of photosynthesis and respiration in two deciduous forests. *Global Change Biology*, 10(6), 939-950.
- Kumar, M. R., Pednekar, S. M., Katsumata, M., Antony, M. K., Kuroda, Y., Unnikrishnan, A. S., 2006. Seasonal variation of the diurnal cycle of rainfall in the eastern equatorial Indian Ocean. *Theoretical and Applied Climatology*, 85(1-2), 117-122.
- Krinner, G., Viovy, N., de Noblet-Ducoudré, N., et al., 2005. A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system. *Global Biogeochemical Cycles*, 19(1).
- Law, B. E., Falge, E., Gu, L.V., et al., 2002. Environmental controls over carbon dioxide and water vapor exchange of terrestrial vegetation. *Agricultural and Forest Meteorology*, 113(1), 97-120.
- Lloyd, J., Taylor, J.A., 1994. On the temperature dependence of soil respiration. *Functional Ecology*, 8, 315-323.
- MÄKELÄ, A., Pulkkinen, M., Kolari, P., et al., 2008. Developing an empirical model of stand GPP with the LUE approach: analysis of eddy covariance data at five contrasting conifer sites in Europe. *Global Change Biology*, 14(1), 92-108.

- Medvigy, D., Wofsy, S. C., Munger, J. W., Hollinger, D. Y., Moorcroft, P. R., 2009. Mechanistic scaling of ecosystem function and dynamics in space and time: Ecosystem Demography model version 2. *Journal of Geophysical Research: Biogeosciences* (2005–2012), 114(G1).
- Menzer, O., Moffat, A. M., Meiring, W., Lasslop, G., Schukat-Talamazzini, E. G., & Reichstein, M., 2013. Random errors in carbon and water vapor fluxes assessed with Gaussian Processes. *Agricultural and forest meteorology*, 178, 161-172.
- Moffat, A. M., Papale, D., Reichstein, M., Hollinger, D. Y., Richardson, A. D., Barr, A. G., ... Stauch, V. J., 2007. Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. *Agricultural and Forest Meteorology*, 147(3), 209-232.
- Moffat, A. M., Beckstein, C., Churkina, G., Mund, M., & Heimann, M., 2010. Characterization of ecosystem responses to climatic controls using artificial neural networks. *Global Change Biology*, 16(10), 2737-2749.
- Moriiasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., Veith, T. L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans. Asabe*, 50(3), 885-900.
- Meyers, S. R., Sageman, B. B., Hinnov, L. A., 2001. Integrated quantitative stratigraphy of the Cenomanian-Turonian Bridge Creek Limestone Member using evolutive harmonic analysis and stratigraphic modeling. *Journal of Sedimentary Research*, 71(4), 628-644.
- Michaelis, L., Menten, M. L., 1913. Die kinetik der invertinwirkung. *Biochem. z*, 49(333-369), 352.
- Milne, B. T., Gupta, V. K., Restrepo, C., 2002. A scale invariant coupling of plants, water, energy, and terrain. *Ecoscience*, 9(2), 191-199.
- Mays, L.W., 2011. *Water Resources Engineering*, second ed. Wiley, Hoboken, NJ.
- Nash, J. E., Sutcliffe, J. V., 1970. River flow forecasting through conceptual models part I -A discussion of principles. *Journal of Hydrology*, 10(3), 282–290.
- Nestler, J. M., Long, K. S., 1997. Development of hydrological indices to aid cumulative impact analysis of riverine wetlands. *Regulated Rivers: Research & Management*, 13(4), 317-334.
- O'Connor, B. L., Hondzo, M., Dobraca, D., LaPara, T. M., Finlay, J.C., Brezonik, P. L., 2006. Quantity-activity relationship of denitrifying bacteria and environmental scaling in streams of a forested watershed. *Journal of Geophysical Research*, 111, G04014.
- Oechel, W.C., Vourlitis, G.L., Verfaillie, J., et al., 2000. A scaling approach for quantifying the net CO₂ flux of the Kuparuk River Basin, Alaska. *Global Change Biology*, 6(S1), 160-173.

- Ooba, M., Hirano, T., Mogami, J. I., Hirata, R., Fujinuma, Y., 2006. Comparisons of gap-filling methods for carbon flux dataset: a combination of a genetic algorithm and an artificial neural network. *Ecological Modelling*, 198(3), 473-486.
- Papale, D., 2012. Data gap filling. In *Eddy Covariance* (pp. 159-172). Springer Netherlands.
- Papale, D., Valentini, R., 2003. A new assessment of European forests carbon exchanges by eddy fluxes and artificial neural network spatialization. *Global Change Biology*, 9(4), 525-535.
- Priestley, M. B., 1981. *Spectral Analysis and Time Series*. Academic Press, London, U.K.
- Raczka, B. M., Davis, K. J., Huntzinger, D., Neilson, R. P., Poulter, B., Richardson, A. D., ... Viovy, N., 2013. Evaluation of continental carbon cycle simulations with North American flux tower observations. *Ecological Monographs*, 83(4), 531-556.
- Richardson, A. D., Hollinger, D. Y., 2005. Statistical modeling of ecosystem respiration using eddy covariance data: maximum likelihood parameter estimation, and Monte Carlo simulation of model and parameter uncertainty, applied to three simple models. *Agricultural and Forest Meteorology*, 131(3), 191-208.
- Reichstein, M., Falge, E., Baldocchi, D., et al., 2005. On the separation of net ecosystem exchange into assimilation and ecosystem respiration: Review and improved algorithm. *Global Change Biology*, 11: 1424–1439.
- Richardson, A. D., Hollinger, D. Y., Aber, J. D., Ollinger, S. V., Braswell, B. H., 2007. Environmental variation is directly responsible for short-but not long-term variation in forest-atmosphere carbon exchange. *Global Change Biology*, 13(4), 788-803.
- Schmidt, A., Wrzesinsky, T., Klemm, O., 2008. Gap filling and quality assessment of CO₂ and water vapour fluxes above an urban area with radial basis function neural networks. *Boundary-Layer Meteorology*, 126(3), 389-413.
- Schmid, H. P., Grimmond, C. S. B., Cropley, F., Offerle, B., Su, H. B., 2000. Measurements of CO₂ and energy fluxes over a mixed hardwood forest in the mid-western United States. *Agricultural and Forest Meteorology*, 103(4), 357-374.
- Schmid, H. P., Su, H. B., Vogel, C. S., Curtis, P. S., 2003. Ecosystem-atmosphere exchange of carbon dioxide over a mixed hardwood forest in northern lower Michigan. *Journal of Geophysical Research: Atmospheres* (1984–2012), 108(D14).
- Schwalm, C. R., Williams, C. A., Schaefer, K., Anderson, R., Arain, M. A., Baker, I., ... Verma, S. B., 2010. A model-data intercomparison of CO₂ exchange across North America: Results from the North American Carbon Program site synthesis. *Journal of Geophysical Research: Biogeosciences* (2005–2012), 115(G3).
- Stauch, V.J., Jarvis, A.J., 2006. A semi-parametric gap-filling model for eddy covariance CO₂ flux time series data. *Global Change Biology*, 12 (9): 1707–1716.

- Stoy, P. C., Dietze, M. C., Richardson, A. D., Vargas, R., Barr, A. G., Anderson, R. S., ... Weng, E., 2013. Evaluating the agreement between measurements and models of net ecosystem exchange at different times and timescales using wavelet coherence: an example using data from the North American Carbon Program Site-Level Interim Synthesis.
- Tian, H., Chen, G., Liu, M., et al., 2010. Model estimates of net primary productivity, evapotranspiration, and water use efficiency in the terrestrial ecosystems of the southern United States during 1895–2007. *Forest Ecology and Management*, 259(7), 1311-1327.
- Thornton, P. E., Law, B. E., Gholz, H. L., et al., 2002. Modeling and measuring the effects of disturbance history and climate on carbon and water budgets in evergreen needleleaf forests. *Agricultural and Forest Meteorology*, 113(1), 185-222.
- Tukey, J. W., 1977 *Exploratory Data Analysis*, Reading, MA: Addison-Wesley
- Urbanski, S., Barford, C., Wofsy, S., Kucharik, C., Pyle, E., Budney, J., ... Munger, J. W., 2007. Factors controlling CO₂ exchange on timescales from hourly to decadal at Harvard Forest. *Journal of Geophysical Research: Biogeosciences* (2005–2012), 112(G2).
- Van Wijk, M. T., Bouten, W., Verstraten, J. M., 2002. Comparison of different modelling strategies for simulating gas exchange of a Douglas-fir forest. *Ecological Modelling*, 158(1), 63-81.
- Valentini, R., Angelis, P. D., Matteucci, G., Monaco, R., Dore, S., Mucnozza, G. S., 1996. Seasonal net carbon dioxide exchange of a beech forest with the atmosphere. *Global Change Biology*, 2(3), 199-207.
- Warnaars, T. A., Hondzo, M., Power M.E., 2007. Abiotic controls on periphyton accrual and metabolism in streams: Scaling by dimensionless numbers. *Water Resources Research*, 43, W08425.
- West, G.B., Brown, J.H., Enquist, B. J., 2001. A general model for ontogenetic growth. *Nature*, 413 (6856), 628-631.
- White, A., Cannell, M.G., Friend, A.D., 1999. Climate change impacts on ecosystems and the terrestrial carbon sink: a new assessment. *Global Environmental Change*, 9, 21-30.
- Wylie, B.K., Fosnight, E.A., Gilmanov, T.G., Frank, A.B., Morgan, J.A., Haferkamp, M.R., Meyers, T.P., 2007. Adaptive data-driven models for estimating carbon fluxes in the Northern Great Plains. *Remote Sensing of Environment*, 106(4): 399-413.
- Xiao, J., Ollinger, S. V., Frohking, S., et al., 2014. Data-driven diagnostics of terrestrial carbon dynamics over North America. *Agricultural and Forest Meteorology*, 197, 142-157.

Appendix

Parameter estimation framework of ESHA

The development of the mathematical framework of ESHA was presented in details by Abdul-Aziz et al. (2007a). Compared to the classical harmonic analysis, we briefly describe the ESHA parameter estimation framework here. According to Priestley (1981), a stochastic Fourier series $y(t)$ is defined as

$$y(t) = a_0 + \sum_{k=1}^W [a_k \cos(2\pi f_k t) + b_k \sin(2\pi f_k t)] + \varepsilon(t) \quad (i)$$

where $y(t)$ = periodic variable to be modeled (e.g., NEE), t = time (independent variable), k = number of harmonics, W = optimum number of harmonics (wave frequencies and corresponding wave patterns are known as harmonics) to avoid frequency aliasing (for proper representation of the diurnal cycle with minimum number of harmonics), a_0 = non-oscillatory ($k = 0$) model component, a_k and b_k = Fourier coefficients associated with the oscillatory components of the harmonic series, $f_k = \frac{k}{t'} = \frac{k}{n\Delta t}$ = k -th harmonic frequency, n = total number of observations within one diurnal cycle t' (e.g., $n = 24$ for $t' = 24$ hours of a diurnal cycle), Δt = sampling interval (e.g., 1 hour), and $\varepsilon(t)$ = random error.

In classical harmonic analysis, all the model parameters (a_0 , a_k and b_k) are estimated simultaneously through a least-squares method by minimizing the sum of the squared errors (

$$SSE = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - h_i)^2)$$
 between the harmonic process $h(t) = y(t) - \varepsilon(t)$ and the

observations of $y(t)$. In contrast, ESHA forces the harmonic process $h(t)$ to pass through a

known value κ (kappa), which is the observation at the reference time (t_{ref}). At that particular reference point (t_{ref}) i.e., $h(t_{ref}) = y(t_{ref}) = \kappa$ (for our application, $\kappa = NEE^*(t_{ref}) = 1.0$) the observed and modeled values are same because $h(t)$ is passed through at the reference point. The non-oscillatory component parameter of a_0 is, therefore, defined from Eq. (1) as

$$a_0 = \kappa - \sum_{k=1}^W [a_k \cos(2\pi f_k t_{ref}) + b_k \sin(2\pi f_k t_{ref})] \quad (ii)$$

The parameters associated with the oscillatory model components (a_k, b_k ; $k = 1, 2, \dots, W$) are estimated first using the least-squares error minimization method. Abdul-Aziz et al. (2007a) presented a system of equations to estimate a_k and b_k as

$$P = Q^{-1}R \quad (iii)$$

where P = parameter (i.e., \hat{a}_k, \hat{b}_k for $k \neq 0$ with ‘hat’ representing estimation) vector of length $2W$, $Q = 2W \times 2W$ non-singular transition matrix, and R = vector of length $2W$ with terms associated with the observations $y(t)$ and the reference-time observation κ . The parameter \hat{a}_0 is then computed from Eq. (ii).

ESHA parameter estimation matrix for two harmonics:

$$\begin{bmatrix} X_{11} & X_{12} & Y_{11} & Y_{12} \\ X_{21} & X_{22} & Y_{21} & Y_{22} \\ U_{11} & U_{12} & V_{11} & V_{12} \\ U_{21} & U_{22} & V_{21} & V_{22} \end{bmatrix} \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n (y_i - \kappa) [\cos(2\pi f_1 t_i) - \cos(2\pi f_1 t_{ref})] \\ \sum_{i=1}^n (y_i - \kappa) [\cos(2\pi f_2 t_i) - \cos(2\pi f_2 t_{ref})] \\ \sum_{i=1}^n (y_i - \kappa) [\sin(2\pi f_1 t_i) - \sin(2\pi f_1 t_{ref})] \\ \sum_{i=1}^n (y_i - \kappa) [\sin(2\pi f_2 t_i) - \sin(2\pi f_2 t_{ref})] \end{bmatrix}$$

where n is the total number of observations in a complete cycle ($n = 24$ for our diurnal NEE cycle), and

$$X_{11} = \frac{n}{2} + n \cos^2(2\pi f_1 t_{ref})$$

$$V_{11} = \frac{n}{2} + n \sin^2(2\pi f_1 t_{ref})$$

$$X_{22} = \frac{n}{2} + n \cos^2(2\pi f_2 t_{ref})$$

$$V_{22} = \frac{n}{2} + n \sin^2(2\pi f_2 t_{ref})$$

$$X_{12} = n \cos(2\pi f_1 t_{ref}) \cos(2\pi f_2 t_{ref})$$

$$V_{12} = n \sin(2\pi f_1 t_{ref}) \sin(2\pi f_2 t_{ref})$$

$$X_{21} = n \cos(2\pi f_2 t_{ref}) \cos(2\pi f_1 t_{ref})$$

$$V_{21} = n \sin(2\pi f_2 t_{ref}) \sin(2\pi f_1 t_{ref})$$

$$Y_{11} = n \cos(2\pi f_1 t_{ref}) \sin(2\pi f_1 t_{ref})$$

$$U_{11} = n \sin(2\pi f_1 t_{ref}) \cos(2\pi f_1 t_{ref})$$

$$Y_{12} = n \cos(2\pi f_1 t_{ref}) \sin(2\pi f_2 t_{ref})$$

$$U_{12} = n \sin(2\pi f_1 t_{ref}) \cos(2\pi f_2 t_{ref})$$

$$Y_{22} = n \cos(2\pi f_2 t_{ref}) \sin(2\pi f_2 t_{ref})$$

$$U_{22} = n \sin(2\pi f_2 t_{ref}) \cos(2\pi f_2 t_{ref})$$

$$Y_{21} = n \cos(2\pi f_2 t_{ref}) \sin(2\pi f_1 t_{ref})$$

$$U_{21} = n \sin(2\pi f_2 t_{ref}) \cos(2\pi f_1 t_{ref})$$

Mathematical expressions for RSR and NSE:

The ration of root-mean-square error to the standard deviation of observations (RSR) can be computed as:

$$RSR = \frac{\sqrt{\frac{\sum_{i=1}^N (NEE_{i,mod} - NEE_{i,obs})^2}{N}}}{\sigma_{NEE_{obs}}}$$

where N = total number of observations, $NEE_{i,obs}$ and $NEE_{i,mod}$ = the i^{th} observed and predicted NEE (respectively), and $\sigma_{NEE_{obs}}$ is the standard deviation of all NEE observations.

The Nash-Sutcliffe Efficiency (NSE) can be calculated as follows:

$$NSE = 1 - \frac{\sum_{i=1}^N (NEE_{i,mod} - NEE_{i,obs})^2}{\sum_{i=1}^N (NEE_{i,obs} - \overline{NEE_{obs}})^2}$$

where $\overline{NEE_{obs}}$ is the average of all NEE observations and others are as previously defined.

Chapter 3: Harmonic modeling of wetland net ecosystem exchange (NEE): Application of scaling for similarity and robust predictions

Abstract

A scaling-based harmonic model is developed by using an extended stochastic harmonic algorithm (ESHA) to achieve robust predictions and simulations of the diurnal cycle of wetland net ecosystem exchange (NEE) from a single reference observation. Scaling of the diurnal NEE cycles was performed by normalizing each diurnal cycle with a corresponding reference-time observation; leading to a single, generalized diurnal curve and dimensionless model of NEE. The scaled, dimensionless model was evaluated with different five year data for a coastal freshwater wetland (Florida Everglades Taylor slough marsh; 2008-12) and an inland shrub wetland (Lost Creek, Wisconsin; 2004-08), as available in the AmeriFlux network. The optimal model involved only five parameters, which exhibited temporal robustness by collapsing into narrow ranges among different days and years. The model parameters did not show any notable trends or patterns with the commonly known drivers (e.g., photosynthetically active radiation, temperature, and wind speed) of wetland NEE. Model calibrations showed impressive efficiency ($R^2 = 0.64$ to 0.70 ; Nash-Sutcliffe Efficiency, $NSE = 0.62$ to 0.66) and accuracy (ratio of root-mean-square-error to the standard deviation of observations, $RSR = 0.58$ to 0.62) in predicting the hourly NEE for the study wetlands. The equally impressive performance was noted in model validations with independent data. Successful predictions of the hourly NEE for different days of different years at each site with the corresponding temporally ensemble set of model parameters indicated the robustness of the scaled model. Model robustness was further demonstrated by analytically deriving the parameter sensitivity and model uncertainty measures, and quantifying them with observed data. The research presents an important example of scaling applications in ecological engineering to achieve ecological similitude and useful generalization in time despite the complex

processes of wetland carbon dynamics. The scaled harmonic NEE model can be used for a robust prediction and gap-filling of continuous (e.g., hourly) NEE time-series of different days and years from the corresponding single (or multiple) reference observations for various wetland ecosystems.

3.1 Introduction

Wetlands are considered one of the most productive ecosystems on earth (Reddy and DeLaune, 2004). Subject to the high rates of carbon sequestration (Bridgham et al., 2006), wetlands play an important role in global carbon dynamics. Wetland net ecosystem exchange (NEE) is a key indicator of wetland carbon balance, representing the net exchange of CO₂ between the ecosystem and the atmosphere (Xiao et al., 2014). Availability of continuous (e.g., hourly) NEE data is essential for accurate estimations of wetland carbon budgets for the relatively longer time-scales (e.g., daily, monthly, annually), which ultimately contributes to the development of strategies for an appropriate management of wetland carbon. NEE generally represents a concave diurnal cycle due to the daytime photosynthesis (downward, negative fluxes) and nocturnal respiration (upward, positive fluxes). However, subject to the influence of diverse biological, hydrological and climatic attributes and drivers (e.g., ecosystem structure, plant type, solar radiation, aerodynamic turbulence, temperature, soil characteristics, water level), the diurnal NEE cycles can substantially vary among different days (both intra- and inter-annually) at the same (Sagerfors et al., 2008) and/or different sites (Moore et al., 2002). Multiple constraints (e.g., instrument failure, rough weather condition, and a breach in micrometeorological hypothesis, lack of resources and manpower) can make it difficult to collect high frequency, continuous NEE for all 24-hours of a day using latest technologies such as the eddy-covariance towers (Baldocchi, 2003) or flux chambers. The sampling constraints result in sparse data or gaps in the measurement of continuous NEE for wetlands. For example, eddy-covariance based AmeriFlux

network (Baldocchi, 2003) encounters on average 35% gaps in their half-hourly observations annually for different ecosystems (Falge et al., 2001). A standard modeling approach is, therefore, needed for robust predictions and gap-filling of continuous (e.g., hourly) NEE for wetland ecosystems from the available sparse or incomplete time-series observations.

Several process-based (i.e., mechanistic) models (e.g., Cao et al., 1996; Potter, 1997; Walter and Heimann, 2000; Zhang et al., 2002; van Huissteden et al., 2006) are available for the estimation and prediction of carbon (CO₂ and CH₄) fluxes mainly from the freshwater wetlands. The wetland-DNDC (Zhang et al., 2002; Cui et al., 2005) exemplifies a commonly used wetland model in the last decade. More recent process models of wetland carbon dynamics can include the LPJ-WHy v1.2 (Wania et al., 2009), McGill (St-Hilaire et al., 2010), ecosys (Dimitrov et al., 2010; Grant et al., 2012), peatland DOS-TEM (Fan et al., 2013), WetQual-C (Sharifi et al., 2013), and CoupModel (Jansson, 2012; Metzger et al., 2015). The mechanistic models are generally built on numerous scientific hypotheses, incorporating different process-components (climate, hydrology, plant biology and ecology, and soil biogeochemistry) and drivers (e.g., vegetation structure and functions, solar radiation, temperature, soil formation, water level) of wetland ecosystems. The process-oriented carbon dynamics models, therefore, often involve a complex structure and are over-parameterized, requiring data for many input variables that are not always available; the large parameter sets and input variables also lead to predictions with high uncertainty.

Data-driven carbon dynamics models can be built on robust empirical patterns hidden in observed data, without necessarily relying on complex scientific hypotheses. Empirical models formulated by mainly considering dominant process variables, as identified by applying ecological data-analytics tools (see Ishtiaq and Abdul-Aziz, 2015), would require a significantly reduced set of input variables and parameters, which can ultimately lead to predictions with much

less uncertainty. The structurally simple empirical models are computationally inexpensive, and can be used as efficient and handy ecological engineering tools for predictions, estimations and gap-filling of wetland NEE without requiring much expert knowledge.

Available data-driven wetland models were developed mainly to explain (rather than predict) the influence of different climatic, hydrologic, ecological and biogeochemical drivers on the wetland carbon (CO₂ and CH₄) fluxes. Previous research (e.g., Yvon-Durocher et al., 2014; Schedlbauer et al., 2012; Yurova et al., 2007; Xing et al., 2005; Smith et al., 2003; Hargreaves and Fowler; 1998; Frolking and Crill, 1994; Dise et al., 1993; Roulet et al., 1992; Moore and Knowles, 1990; Bartlett et al., 1987; Sebacher et al., 1986) presented regression-based empirical models mainly for explaining the correlations of carbon fluxes with the climate, soil, and other environmental variables. Furthermore, various empirical models of CO₂ fluxes were used to fill gaps in the relatively large, continuous (e.g., half-hourly) database of the AmeriFlux network, which includes many terrestrial (and some wetland) ecosystems. These empirical gap-filling methods include the application of mean diurnal variation (Falge et al., 2001); ‘look-up’ tables developed based on ambient environmental conditions (Falge et al., 2001); and parametric and non-parametric statistical approaches, including artificial neural networks and non-linear regressions (e.g., Stauch and Jarvis, 2006; Braswell et al., 2005; Jarvis et al., 2004; Hui et al., 2004; Papale and Valentini, 2003; Aubinet et al., 1999; Valentini et al., 1996; Lloyd and Taylor, 1994; Michaelis and Menten, 1913).

The cyclic pattern of ecological and environmental variables such as the wetland NEE inspires the development of relatively simple, predictive harmonic (i.e., empirical) models by leveraging the well-known Fourier series expansion (Priestley, 1981). A remarkable advantage of harmonic analysis over other regression (empirical) or complex process-based models is that the Fourier model development mainly requires data for only the predictant (e.g., NEE) itself. An

optimal (parsimonious) harmonic model can also provide useful predictions with a small set of parameters, which can lead to reduced model uncertainty (e.g., see Abdul-Aziz et al., 2007; Abdul-Aziz and Ishtiaq, 2014). Although the classical harmonic analysis (Priestley, 1981) has been widely used to develop empirical models in numerous applications of environmental and water resources engineering (e.g., Kumar et al., 2006; Dyar and Alhadeff, 2005; Meyers et al., 2001; and Nestler and Long, 1997), only a few studies applied harmonic functions for the modeling of ecosystem CO₂ exchanges. Richardson and Hollinger (2005) and Hollinger et al. (2004) developed second order harmonic regression models to predict nocturnal CO₂ fluxes of terrestrial ecosystems. Griffis and Rouse (2001) utilized the fast Fourier transformation and power spectral analysis to predict the inter-annual variability of NEE from a northern wetland (i.e., peatland).

A general shortcoming of both the empirical and process-oriented carbon dynamics models is that the model parameters are mostly specific to the particular sites and/or time-frames of evaluation data, often requiring a new model calibration for a different site or time domain. Application of scaling can transform observations of different time and/or space domains to a comparable reference platform. Scaling has been widely used in numerous environmental science and water resources engineering applications (e.g., Hondzo and Wannaars, 2008; Wannaars et al., 2007; Milne et al., 2002; and O'Connor et al., 2006) to develop robust models by leveraging the underlying similitude (similarity) among different references. Abdul-Aziz et al. (2007a) proposed a novel scaling approach to develop an extended stochastic harmonic analysis algorithm (ESHA), which can convert periodic variables (e.g., stream dissolved oxygen, DO) measured at different clock-times to that at a referenced time (e.g., noon) of the day. Abdul-Aziz and Ishtiaq (2014) expanded the ESHA to develop a scaling-based empirical model that can robustly predict the entire diurnal cycles of stream DO for different days and sites from the corresponding single reference observations. Following the successes, Ishtiaq and Abdul-Aziz (2015) leveraged the

scaling and ESHA framework to successfully develop a spatiotemporally robust model for the prediction and gap-filling of terrestrial NEE time-series with a high resolution (e.g., hourly) based on limited available data by considering five different AmeriFlux deciduous forests as the case studies (see chapter 3). Upscaling of the five forest-specific models to a generalized deciduous forest model (i.e., one set of parameter values for all forests, seasons, and years) showed a remarkable robustness in predicting the diurnal NEE cycles of different deciduous forest sites with impressive model efficiency (e.g., coefficient of determination, $R^2 = 0.60$ to 0.81) and accuracy (ratio of root-mean-square error to the observed standard deviation, $RSR = 0.46$ to 0.64).

This paper leverages the ESHA-based stream DO model of Abdul-Aziz and Ishtiaq (2014) (see chapter 2) to develop a scaling-based harmonic model to robustly predict/simulate the different diurnal cycles of wetland NEE from the corresponding single reference observations. The study builds on a hypothesis that an appropriate scaling of the different diurnal cycles of NEE by the corresponding single reference observations would lead to a general, dimensionless diurnal NEE curve for the same or similar wetlands. The generalized NEE curve, once fitted with observations, would provide a parsimonious set of parameters that are relatively invariant (i.e., stationary) in time, leading to an empirical model to robustly predict wetland NEE with high temporal resolutions. The model is evaluated with hourly NEE observations of different days for different years from two AmeriFlux wetland sites, representing a coastal freshwater wetland and an inland shrub wetland (i.e., to incorporate the coastal vs. inland gradient). Analytical measures of model sensitivity and uncertainty are derived and quantified with the evaluation data to further demonstrate the model robustness.

3.2 Materials and methods

3.2.1 The scaling-based harmonic model theory

The underlying theory of the proposed harmonic model is that wetland NEE follow cyclic (periodic) diurnal patterns, which can be generalized into a single, dimensionless NEE diurnal cycle through scaling (normalizing) by the corresponding single reference observations. The assumption of periodic diurnal variation for NEE appears reasonable year round for tropical wetlands, while representing mainly the growing seasons (e.g., June-September) for wetlands of the temperate and colder regions. Abdul-Aziz and Ishtiaq (2014) described the ESHA modeling concept and scaling application for stream DO. Following their work, a conceptual illustration is made by using a schematic of several hypothetical diurnal cycles of NEE representing different days (1, 2, ..., N) for a single or similar wetland sites (Figure 2.2a). An un-scaled harmonic model would ideally require N unique sets of parameter values to represent the N diurnal cycles of NEE. In contrast, the scaling-based ESHA (Abdul-Aziz et al., 2007a; Abdul-Aziz and Ishtiaq, 2014) can provide a single set of model parameters across time. The ESHA scales (i.e., normalizes) each diurnal cycle by a corresponding reference-time (t_{ref}) single observation NEE_{ref} . The normalization leads to a desired collapse of different NEE cycles on a generalized, dimensionless diurnal NEE curve (NEE^*) that has a value of unity (i.e., $NEE^* = 1.0$) at the t_{ref} (Figure 2.2b). The scaled model (NEE^*) is then estimated with the ESHA parameterization framework (Section 4.2.2) using observational data. The dimensionless model is ideally robust because it represents all simulation days at the same or similar sites with a single set (instead of N sets for N days) of parameters. The different day-specific NEE cycles are then obtained by multiplying the estimated NEE^* model by the corresponding single reference observations, NEE_{ref} .

3.2.2 Estimation of ESHA model parameters

The theoretical framework of ESHA was originally developed and presented in rigorous details by Abdul-Aziz et al. (2007a). A simplified summary of the ESHA parameter estimation framework was also presented by Abdul-Aziz and Ishtiaq (2014). To aid the readers, the mathematical framework of ESHA is also briefly, but adequately described in this paper. A stochastic Fourier series, $y(t)$ (Priestley, 1981) is defined as

$$y(t) = a_0 + \sum_{k=1}^W [a_k \cos(2\pi f_k t) + b_k \sin(2\pi f_k t)] + \varepsilon(t) = h(t) + \varepsilon(t) \quad (1)$$

where $y(t)$ = periodic response variable (i.e., NEE^*), t = time (predictor variable), k = number of harmonics, W = optimum number of harmonics for avoiding frequency aliasing, a_0 = non-oscillatory ($k = 0$) model component, a_k and b_k = Fourier coefficients associated with the oscillatory components of the harmonic series, $f_k = \frac{k}{t'} = \frac{k}{n\Delta t} = k^{\text{th}}$ harmonic frequency, n = total number of observations within one diurnal cycle t' (e.g., $n = 24$ for $t' = 24$ hours of a diurnal cycle), Δt = sampling interval (e.g., 1 hour), $\varepsilon(t)$ = random error, and $h(t) = y(t) - \varepsilon(t)$ is the harmonic process.

Unlike classical harmonic analysis to simultaneously estimate the model parameters (a_0 , a_k and b_k) by a least-squares error minimization between the model and observations, the ESHA first forces the harmonic process $h(t) = y(t) - \varepsilon(t)$ to pass through a known value κ (kappa), which is the observation at the reference time (t_{ref}). This additional forcing results in a “zero” modeling error at t_{ref} , i.e., $h(t_{ref}) = y(t_{ref}) = \kappa$; based on Fig. 1b, $\kappa = NEE^*(t_{ref}) = 1.0$

for our NEE modeling. The non-oscillatory model component parameter of a_0 is then derived from Eq. (1) as

$$a_0 = \kappa - \sum_{k=1}^W [a_k \cos(2\pi f_k t_{ref}) + b_k \sin(2\pi f_k t_{ref})] \quad (2)$$

First, the parameters associated with the oscillatory model components ($a_k, b_k; k = 1, 2, \dots, W$) are estimated using a least-squares error minimization method from a system of equations as (Abdul-Aziz et al., 2007a)

$$P = Q^{-1}R \quad (3)$$

where P = parameter (i.e., \hat{a}_k, \hat{b}_k for $k \neq 0$ with ‘hat’ representing estimation) vector of length $2W$, $Q = 2W \times 2W$ non-singular transition matrix, and R = vector of length $2W$ with terms associated with the observations $y(t)$ and the reference-time observation κ . The parameter \hat{a}_0 is then computed from Eq. (2). An explicit form of Eq. (3) for two harmonics (i.e., $W = 2$) is presented in the Appendix of chapter 3 for convenience of the readers.

Compared to a classical harmonic analysis, the ESHA provides the correct parameter estimation framework to force the harmonic model through the reference-time normalized observation (e.g., $NEE^*(t_{ref}) = 1.0$). The forcing, in turns, allows the proper reconstruction of the entire diurnal cycle of NEE from a single reference observation.

3.2.3 The ESHA modeling algorithm for wetland NEE

Development of the scaling-based ESHA model requires continuous (e.g., hourly) NEE data for at least one (and for preferably multiple) 24-hour diurnal cycle(s) for a single or multiple similar

wetlands. Model application requires at least a single reference-time observation to predict the corresponding diurnal cycle. The modeling algorithm is stated as follows:

1. Collect hourly (t) NEE data, $NEE(t)$ for multiple 24-hour cycles.
2. Select a reference-time (t_{ref}) observation, $NEE_{ref} = NEE(t_{ref})$ for each cycle.
3. Normalize (i.e., scale) each NEE cycle by the corresponding day-specific reference observation (NEE_{ref}) to obtain a dimensionless cycle as: $NEE^*(t) = NEE(t) / NEE_{ref}$.
4. Fit each diurnal cycle of $NEE^*(t)$ by applying the ESHA parameter estimation framework (Section 4.2.2).
5. Obtain a robust (generalized for different days) scaled model, $NEE_{mod}^*(t)$ with the ensemble averages of daily estimated parameters.
6. Predict the individual diurnal cycles, $NEE_{mod}(t)$ by multiplying the generalized scaled model, $NEE_{mod}^*(t)$ by the corresponding single reference observations (NEE_{ref}) as:

$$NEE_{mod}(t) = NEE_{ref} \times NEE_{mod}^*(t).$$

3.2.4 Study wetlands and data sets

Although the eddy-covariance technique has become a major mode of measuring ecosystem level CO_2 fluxes, wetlands are not extensively represented in the flux (i.e., AmeriFlux) network (Lund et al., 2010). In fact, AmeriFlux (Baldocchi, 2001) includes a limited number of wetlands and all of them are not fully operational. Choice of the wetland case study sites was primarily based on the availability of continuous (e.g., hourly) NEE data with periodic diurnal patterns. Further, the presence of spatial, climatic, hydrological, biogeochemical, and ecological gradients among different study sites were considered critical to test the model effectiveness for different wetland

types. Based on the criteria, the scaling-based harmonic NEE model were evaluated for two wetlands (Table 3.1, Figure 3.1) from the AmeriFlux network (AmeriFlux website, 2014): (1) Florida Everglades Taylor Slough short hydroperiod marsh, Florida (US-Esm) and (2) Lost Creek, Wisconsin (US-Los); representing coastal freshwater marsh and inland shrub wetland, respectively

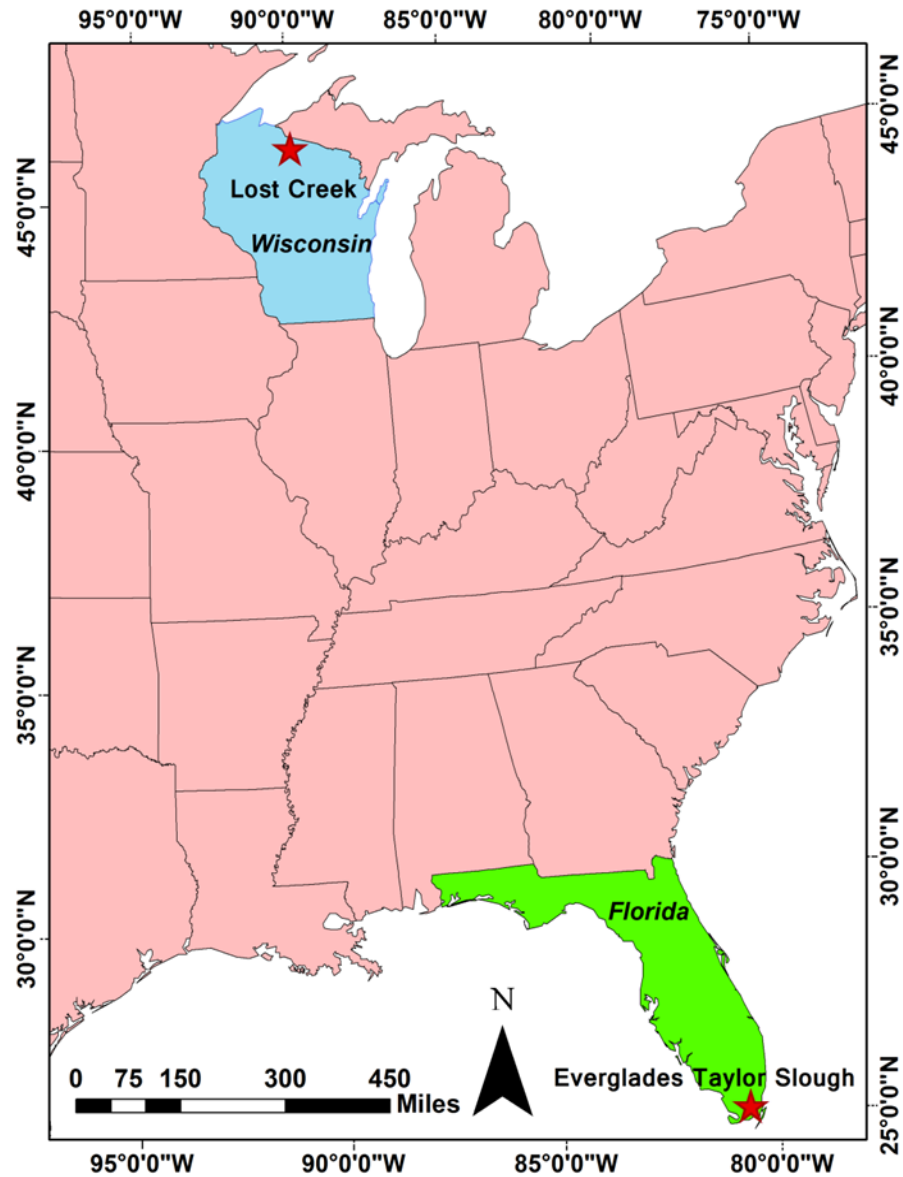


Figure 3.1: Study area and geographic locations of the selected AmeriFlux wetland study sites

The Taylor Slough marsh typically experiences seasonal flooding for approximately five months in a year (Schedlbauer et al., 2010). It has been one of the extensively managed wetlands in terms of hydrology, with water flows being controlled by a combination of optimized pumping and constructed levees (Kotun and Renshaw, 2014). Sawgrass (*Cladium jamaicense*) and C4 muhly grass (*Muhlenbergia capillaris*) are the most prevailing vegetation and primary producers at this site; additionally, periphyton and submerged macrophytes geochemically fixed CO₂ (e.g., as CaCO₃) during flooding seasons (Schedlbauer et al., 2012). The Lost Creek is a minerotrophic shrub wetland located around 480 m above the mean sea level (Sulman et al., 2009). Alder (*Alnus incana subsp. rugosa*), willow (*Salix sp.*) and sedges (*Carex sp.*) dominate the vegetation around the flux tower at this site. Soil is poorly drained, mostly composed of Totagatic-Bowstring-Ausable complex, Seelyeville and Markey mucks.

Table 3.1: Summary of the study sites and NEE data collection period

Site name	Latitude (°N)	Longitude (°W)	Climate	Data Period
Florida Everglades Taylor Slough Short Hydroperiod Marsh, FL (US-Esm)	25.44	80.59	Tropical	2008-12
Lost Creek, WI (US-Los)	46.08	89.98	Northern Continental	2004-08

The hourly NEE ($\mu\text{mol}/\text{m}^2/\text{s}$) observations of different 24-hour (starting from 1:00 A.M.) diurnal cycles during 2004-12 (US-Los: 2004-08, and US-Esm: 2008-12) were sub-sampled from the AmeriFlux half-hourly time-series. In addition, daily average photosynthetic active radiation (PAR), air temperature (TA), wind speed (WS), and water level (WL) were obtained from the AmeriFlux network to investigate the temporal relationship between the model parameters and the important drivers of wetland NEE at the study sites. The selected data sets would reflect the impact of seasonal and inter-annual variability of the wetland carbon fluxes and associated major drivers in evaluating the model performance.

Based on the standard AmeriFlux sign conventions, negative NEE indicates the downward (atmosphere to wetland) fluxes and positive NEE represents the upward (emissions to the atmosphere) fluxes. The hourly NEE time-series of the study periods showed year-round productivity for the Taylor Slough (US-Esm); the Lost Creek (US-Los) was substantially less productive during winter and spring compared to that for the growing seasons (Figure 3.2). Taylor Slough NEE also showed clear diurnal cycles for the entire year, whereas a periodic diurnal pattern for the Lost Creek were apparent mainly during the growing period (June-September). Since ESHA assumes periodic diurnal cycles in NEE, hourly gap-filled data for four years (2008-11) (total 1461 diurnal cycles) were considered for parameter estimation and model calibration at the Everglades Taylor Slough site. In contrast, the extended growing period (June-September) NEE time-series between 2005 and 2008 (480 diurnal cycles) were used for model calibration at the Lost Creek wetland site. However, un-filled NEE data of different one year periods (independent of the model calibration years; Taylor Slough: 2012, Lost Creek: 2004) were used for the model validations. Consideration of un-filled data for model validations would strengthen the model evaluation process by avoiding any bias originating from the usage of gap-filled data during the calibration phase of the model.

3.3 Results

3.3.1 Selection of an optimal model

Optimal number of harmonics

Following Abdul-Aziz and Ishtiaq (2014), a slightly modified version of the Akaike Information Criterion (AIC) (Akaike, 1974) was used to determine the optimal number of harmonics for a parsimonious model (involving minimum parameters) without compromising the prediction accuracy. For both the study sites, two harmonics ($W = 2$) resulted in the minimum AIC to

optimally predict hourly NEE^* by the dimensionless, scaled model over the entire study periods (Figure 3.3a). The selection limited the number of model parameters to five dimensionless numbers (\hat{a}_0 , \hat{a}_1 , \hat{a}_2 , \hat{b}_1 , and \hat{b}_2) for each site. Ishtiaq and Abdul-Aziz (2014) and Abdul-Aziz et al. (2007) also considered an optimal number of two harmonics for their scaling based ESHA models to predict terrestrial NEE and stream DO, respectively.

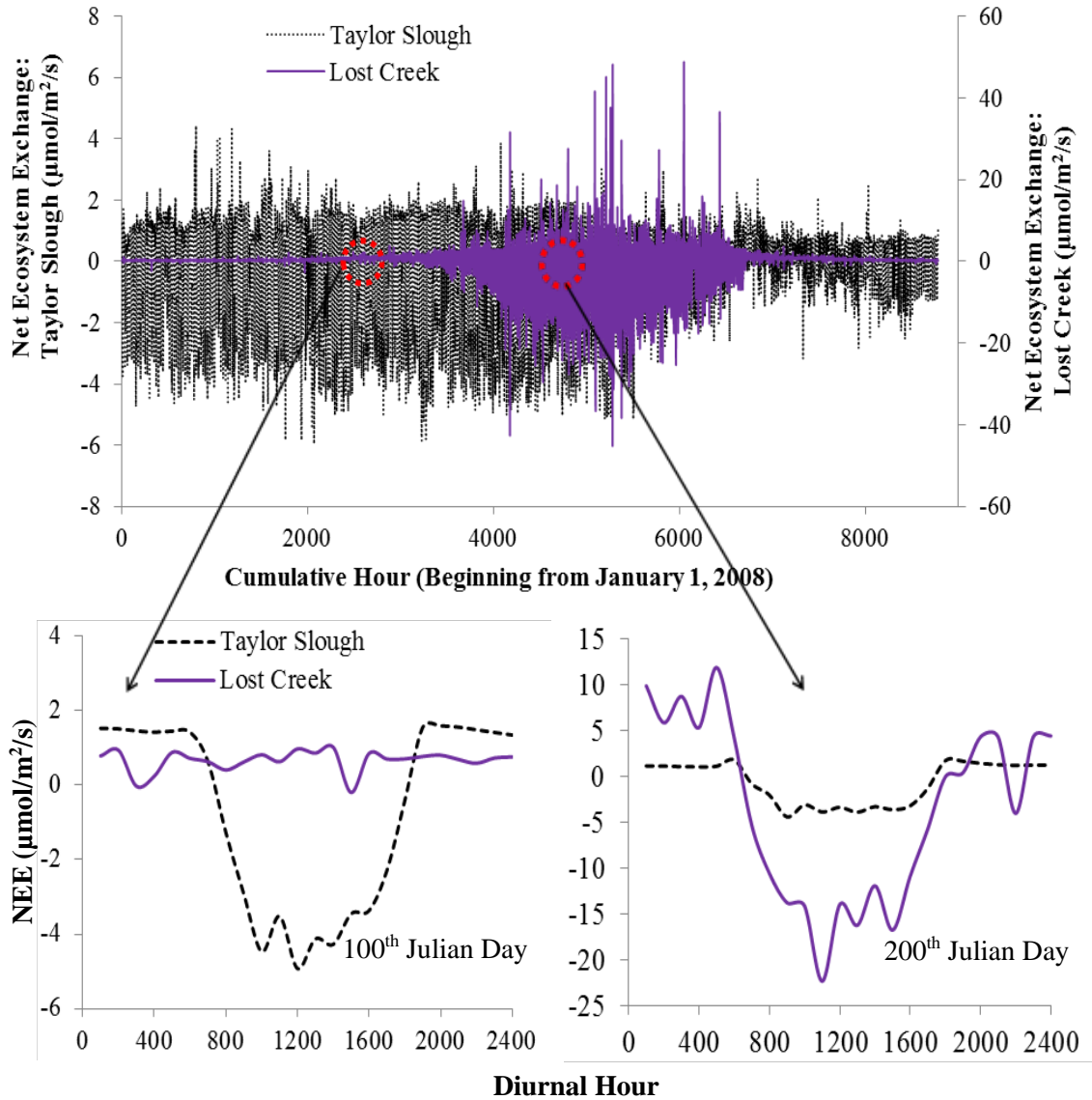


Figure 3.2: Plot of the hourly net ecosystem exchange (NEE) of Taylor Slough and Lost Creek sited for 2008 showing variation in diurnal cycles of NEE (upper panel). 100th and 200th Julian days are zoomed in for Taylor slough and Lost creek from 2008 hourly NEE data to see the diurnal variation of NEE for these two sites at different season (lower panel).

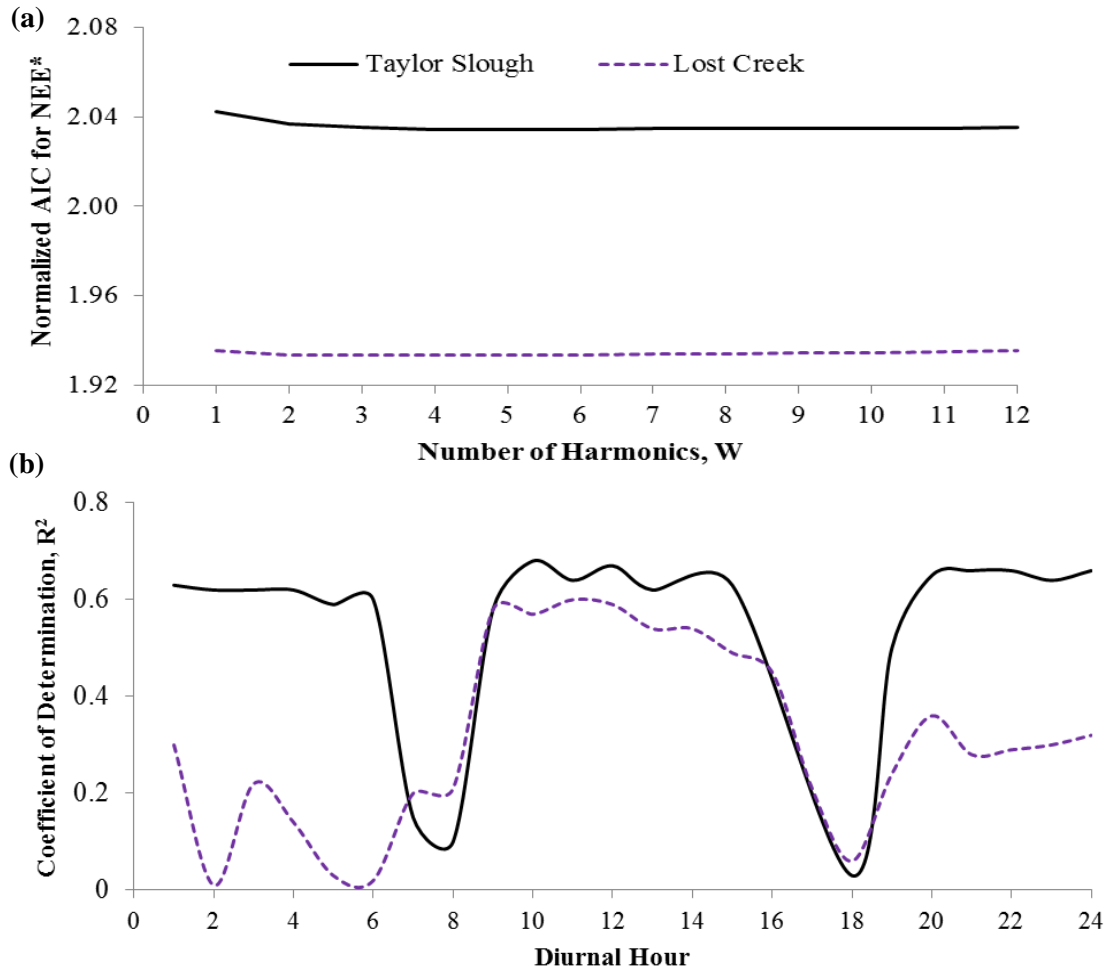


Figure 3.3: Optimization of the ESHA model for the selection of (a) maximum number of harmonics required for model simulation by plotting the normalized AIC_{NEE^*} with the number of harmonics (W) and (b) best reference time (t_{ref}) from the 24-hour period that could be used for modeling from a plot of coefficient of determination, R^2 and different diurnal hours for two study wetlands.

Optimal reference time

In theory the ESHA framework allows the selection of any diurnal hour as the reference-time (t_{ref}) and the associated NEE observation as the reference observation for scaling the respective diurnal cycle. The impact of using different diurnal hours (from the 24-hour cycle) as the reference time on the wetland NEE^* model predictions was assessed by plotting the chosen diurnal hours versus the coefficient of determination (R^2), as obtained by using a two harmonic

model. As shown in Figure 3.3b, the time-window of 10-14 hours (i.e., 10:00 A.M. to 2:00 P.M.) provided the higher model prediction efficiencies (i.e., explaining more variance in observed data) for both study sites. Specifically, it is reasonable to state that scaling different diurnal cycles by the corresponding midday (12:00 P.M.) reference observations provided an optimal modeling performance for both the Taylor Slough (coastal Everglades freshwater wetland) and Lost Creek (inland shrub wetland) study sites. Similar finding was reported by previous studies (e.g., Abdul-Aziz et al., 2007a; Abdul-Aziz and Ishtiaq, 2014) that predicted the different diurnal cycles of stream DO using the ESHA framework.

3.3.2 Robust estimation of model parameters

The site-specific model fitting and calibration estimated five dimensionless model parameters (\hat{a}_0 , \hat{a}_1 , \hat{a}_2 , \hat{b}_1 , and \hat{b}_2) of the scaled NEE model (NEE_{mod}^*) with hourly NEE data of 1461 days (January-December, 2008-11) and 480 days (June to September, 2005-08) from, respectively, the Taylor Slough and Lost Creek wetlands. To avoid model bias toward any outlying diurnal cycles for each site, the respective daily parameter values outside their 1st and 99th percentiles were identified, and the corresponding diurnal cycles of observed NEE were removed from the calibration data sets for both study sites. This filtering led to the exclusion of around 4% and 15% of the NEE data-days for Taylor Slough and Lost Creek sites, respectively. The model parameters were re-estimated by using the filtered data sets for Taylor Slough (1403 diurnal cycles) and Lost Creek (408 cycles). The 4-year ensemble sets of different parameters for each wetland site were obtained by averaging the respective daily estimates (Table 3.2).

The multi-year, multi-season temporal means of constant parameter \hat{a}_0 (the non-oscillating, major model component representing a quasi-average daily NEE*) were -0.053 and 0.136 for US-Esm and US-Los, respectively. Observed significant difference of \hat{a}_0 between the

wetland sites revealed two separate climatic and biogeochemical processes that lead to a variability in NEE, which is also apparent from Figure 3.2. However, first harmonic parameters of \hat{a}_1 and \hat{b}_1 (representing the primary oscillations of the diurnal cycle) and the second harmonic parameters of \hat{a}_2 and \hat{b}_2 (representing the secondary oscillations) had more comparable values for the two drastically different case study wetlands (Table 3.2).

Table 3.2: Means of (averaged over the calibration period) estimated model parameters at study sites

Site name	\hat{a}_0	\hat{a}_1	\hat{a}_2	\hat{b}_1	\hat{b}_2
Taylor Slough (US-Esm), FL	-0.053 (0.409)	-0.857 (0.411)	0.196 (0.186)	-0.091 (0.20)	-0.050 (0.191)
Lost Creek (US-Los), WI	0.136 (0.322)	-0.756 (0.535)	0.109 (0.419)	0.102 (0.352)	-0.058 (0.320)

Notes: Values in parentheses are the standard deviations of the estimated parameters. Hat (^) sign represents fitted or estimated parameters of the Fourier series.

Temporal trends of model parameters were investigated by plotting the filtered (i.e., within the respective 1st and 99th percentiles), day-specific estimates of \hat{a}_0 , \hat{a}_1 and \hat{b}_1 with the corresponding cumulative parameterization days of calibrations for the study wetlands (Figure 3.4a,b). The parameters (\hat{a}_0 , \hat{a}_1 and \hat{b}_1) did not show any notable increasing or decreasing trend over the estimation period revealing temporal robustness of the model parameters. However, both \hat{a}_0 and \hat{a}_1 showed clustered dispersions for each year in late October (late fall) at US-Esm (Fig 3.4a); this is likely caused by the less well-defined diurnal cycles of NEE during transitions with the winter season. Dispersion in parameter values at the end of September (Fig 3.4b) was observed in the US-Los. However, consideration of higher percentile values during filtering could eliminate these dispersions, which would reduce the calibration data set of US-Los. Overall, the model parameters collapsed into narrow bands among different days for both the wetlands, demonstrating impressive temporal robustness.

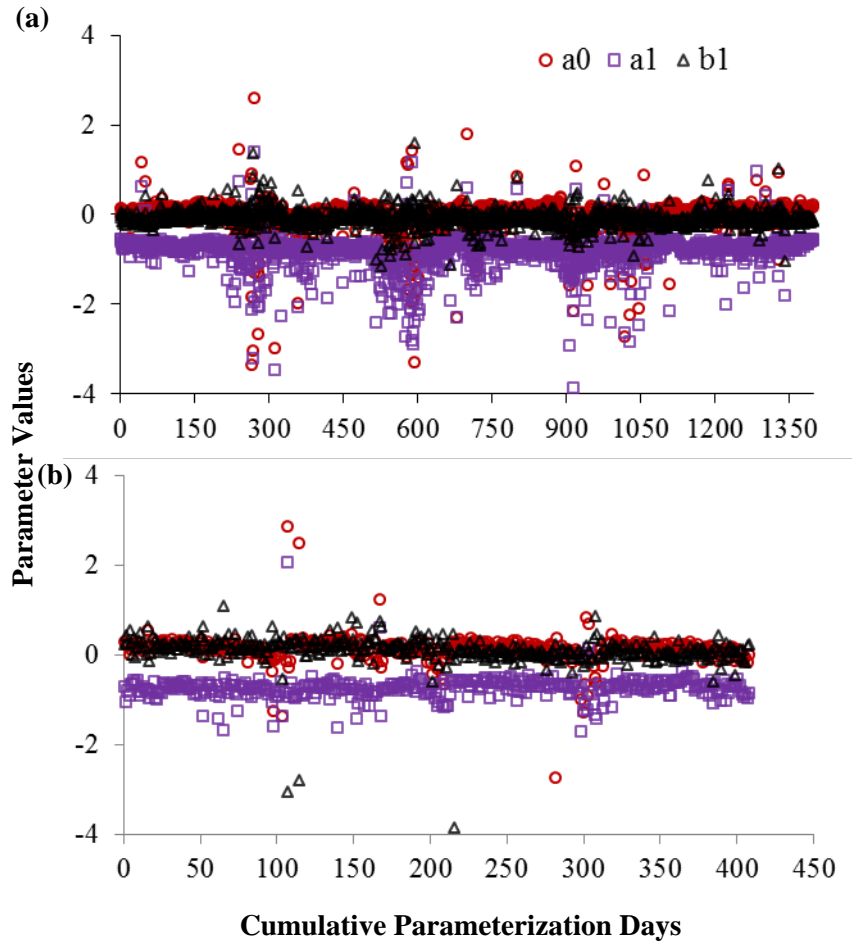


Figure 3.4: Daily time-series of estimated parameters during the four years of calibration for the study wetlands, demonstrating the temporal robustness of the model parameters. The a_0 , a_1 , and b_1 , respectively, refer to \hat{a}_0 , \hat{a}_1 and \hat{b}_1 where the hat ($\hat{\cdot}$) represents estimations.

The relationships between the model parameters and the major process drivers of NEE were investigated for the study wetlands to attain mechanistic insights into the temporal robustness of estimated parameters. As an example, the daily estimates of the principal parameter, \hat{a}_0 and the 1st harmonic parameters (\hat{a}_1 and \hat{b}_1) were plotted with the corresponding 24-hour average air temperature (TA), wind speed (WS) and photosynthetic active radiation (PAR) of 2008 for the study sites for (see Figure 3.5). The daily TA, WS and PAR were normalized by their corresponding temporal averages (annual average for US-Esm and June-September average

for US-Los) to bring them on a comparable reference scale. Plotted parameters did not display any noteworthy trend with the TA, WS and PAR for the wetlands, reemphasizing the robustness of model parameters. However, the higher dispersion of \hat{a}_0 and \hat{a}_1 for the US-Esm suggested more variable response of the quasi-average dimensionless daily NEE to the major process drivers at this site than that for the other study wetland; this can also explain the higher dispersions of \hat{a}_0 and \hat{a}_1 when plotted with the cumulative model days (see Figure 3.4a).

The observed collapse (i.e., robustness) of the estimated model parameters is a unique outcome of scaling; normalization of the each diurnal NEE cycle by the corresponding single reference observation brought parameters of different days into the comparable ranges. The impressive temporal robustness of the model parameters suggest that hourly NEE predictions made by using the aggregations (i.e., site-specific ensemble mean) of daily parameters over the 4-year calibration periods would also be robust in time.

3.3.3 Model evaluations and robust predictions

The site-specific, temporally ensemble parameters (Table 3.2) were used to predict the hourly NEE cycles from the corresponding day-specific single reference observations for all the filtered days (US-Esm: 1403 days; US-Los: 408 days) of the study wetlands. Three criteria were used to evaluate the model performance (Table 3.3): (i) ratio of root-mean-square error to the standard deviation of observations (RSR) that indicates model accuracy (Singh et al., 2004); (ii) coefficient of determination (R^2), indicating the amount of observed data variance explained by the model ($R^2=1.0$ indicates a perfect model); and (iii) Nash-Sutcliffe Efficiency (NSE) (Nash and Sutcliffe, 1970) that also measures the goodness of model fit (see supplementary note for the mathematical expressions of RSR and NSE). Following the extensive analysis of Moriasi et al. (2007) on model performance, an RSR from 0 to 0.50 indicates a perfect to very good model, from 0.5 to 0.6 indicates a good model, and from 0.6 to 0.7 refers to a satisfactory model; a model with RSR >

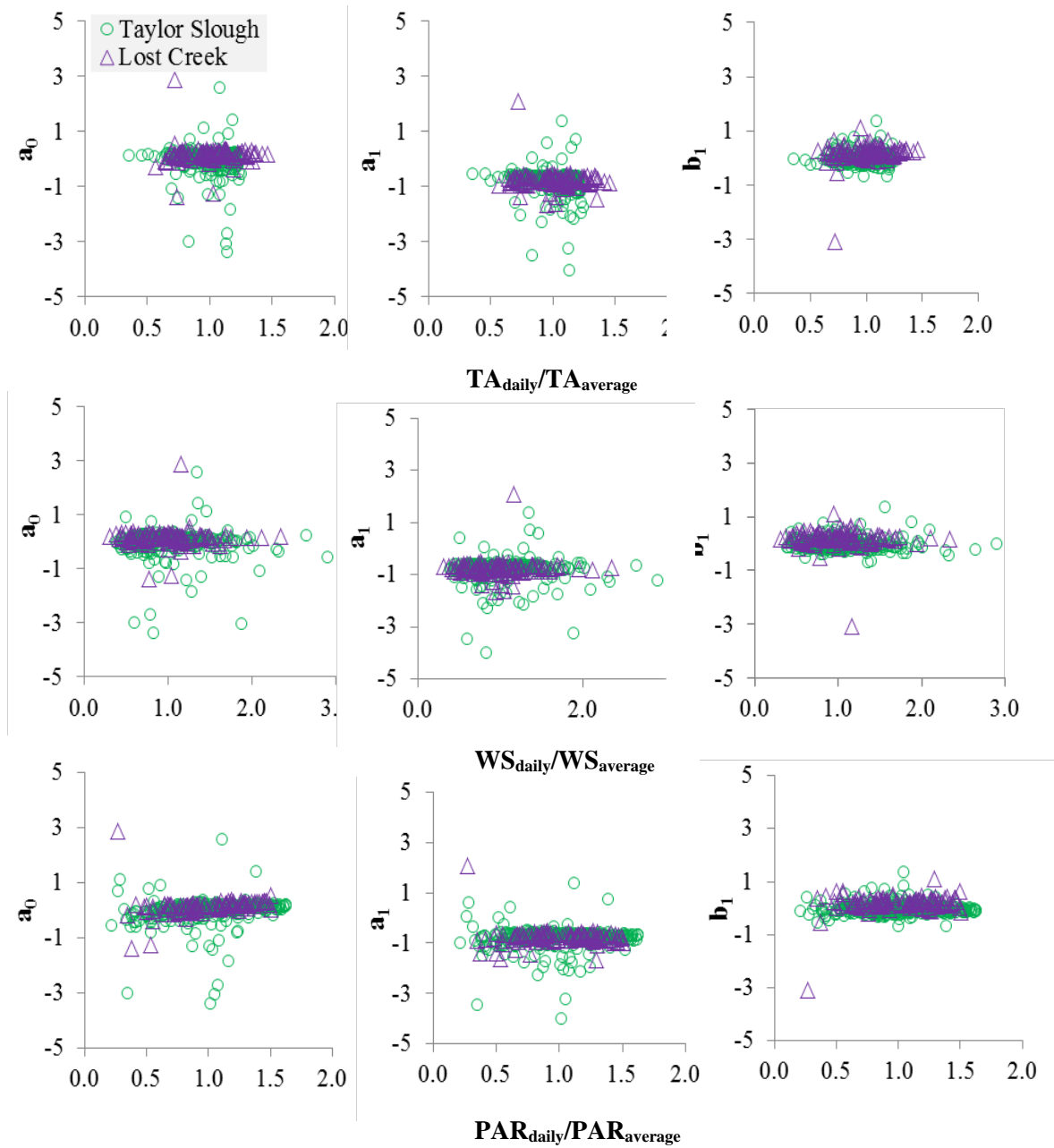


Figure 3.5: Plots of the daily estimates of the principal model parameter, \hat{a}_0 and 1st harmonic parameters \hat{a}_1 and \hat{b}_1 with the corresponding (as normalized by the respective site-specific averages), air temperature ($TA_{\text{daily}}/TA_{\text{average}}$), wind speed ($WS_{\text{daily}}/WS_{\text{average}}$) and photosynthetic active radiation ($PAR_{\text{daily}}/PAR_{\text{average}}$) for the year 2008 at two wetland ecosystems. The a_0 , a_1 , and b_1 , respectively, refer to \hat{a}_0 , \hat{a}_1 and \hat{b}_1 where the hat (^) represents estimations.

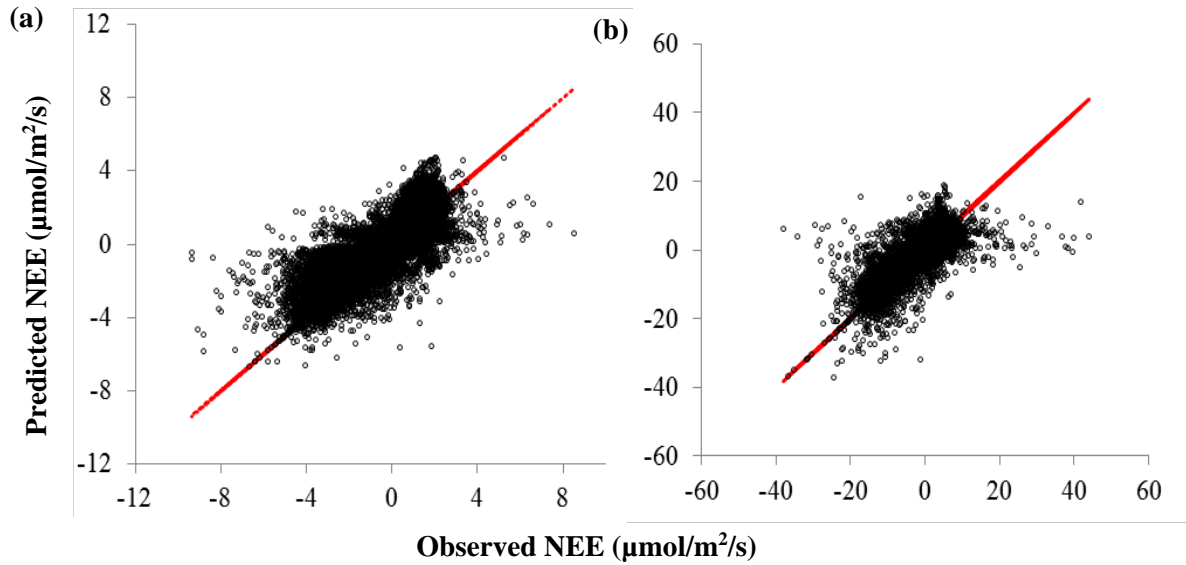


Figure 3.6: Calibration plots of the hourly predicted and observed NEE of different calibration periods for the (a) Taylor Slough (US-Esm) and, (b) Lost Creek (US-Los) using temporally ensemble means of site-specific model parameters. Dashed red line represents the 1:1 perfect fit line.

Table 3.3: Summary of the results for site specific model calibration and validation at different study sites

Site	Year	Calibration			Validation			
		R ²	RSR	NSE	Year	R ²	RSR	NSE
Taylor Slough (US-Esm)	2008-11	0.70	0.58	0.66	2012	0.70	0.58	0.66
Lost Creek (US-Los)*	2005-08	0.64	0.62	0.62	2004	0.63	0.64	0.59

Note: At US-Los hourly NEE data from June to September of different calibration and validation years were used

0.70 is considered unsatisfactory. Similar to the R², the NSE of 1.0 indicates a perfect model, whereas NSE > 0 indicates better fitting with the harmonic model than that of the daily mean of hourly observations as an alternative model. Model calibration showed an impressive model fitting (R²: 0.70 and 0.64; NSE: 0.66 and 0.62) and accuracy (RSR: 0.58 and 0.62) in US-Esm and US-Los, respectively (Table 3.3, Figure 3.6). However, as shown in Figure 3.6, model was not able to predict some of the extreme values of NEE properly; the long-term temporal aggregation of model parameters might have caused this extreme value induced inaccuracy.

Although relatively larger data set was considered for model calibration, US-Esm demonstrated better model fitting compare to US-Los.

The calibrated NEE model was validated with independent single year of NEE data of 2012 and 2004 (June- September) for US-Esm and US-Los, respectively (Table 3.3, Figure 3.7). The data that were used for validation was not gap filled to test the ESHA modeling efficiency for both gap-filled and non-gap -filled data as calibration was done from gap-filled data. Since day-specific single NEE observation at 12: PM (NEE_{ref}) is required to multiply the NEE_{mod}^* (estimated from the calibration parameters) (Equation 4) for the prediction of the entire diurnal cycle during validation period, the days having observed NEE at 12:00 PM were considered for validation. As such 191 and 89 days of NEE diurnal cycles (outside of the respective calibration years) were predicted from their corresponding reference-time single observations by using the previously estimated site-specific, temporally ensemble parameters. The model validation demonstrated nearly equivalent prediction performance (RSR = 0.58 and 0.64; $R^2 = 0.70$ and 0.63; and NSE = 0.66 to 0.59), compared to that of the calibrations among the study sites (Figure 3.8).

Overall, the site-specific model calibrations and validations revealed an impressive ability of the scaling-based ESHA model in predicting the entire diurnal cycles of hourly NEE from the respective single reference observations. The success in simulating hourly NEE by using the temporally ensemble (multi-year, multi-season means) single set of site-specific model parameters, also showed the temporal robustness of the scaled model. Since the diurnal NEE cycle varies significantly among different days at the same site, a harmonic model without scaling cannot predict the different diurnal NEE cycles from a set of ensemble parameters. Scaling by the day-specific single reference observations resulted in the collapse of different cycles into a general dimensionless diurnal curve (as shown in Figure 2.2) and provided the comparable values

of estimated parameters over the entire simulation periods; which led to the temporally robust predictions for each wetland

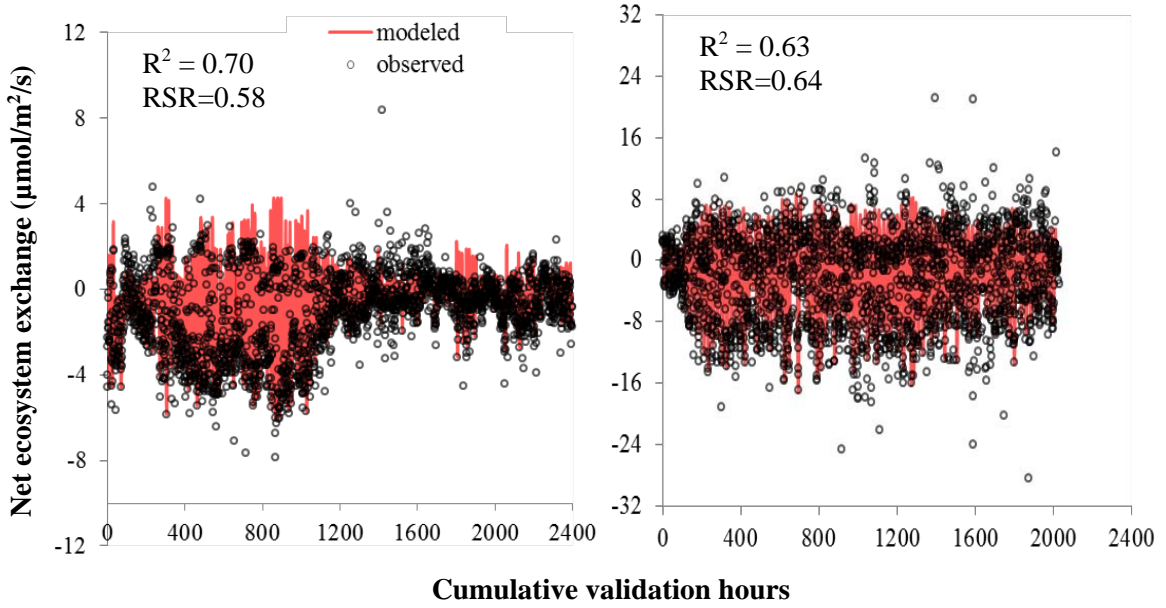


Figure 3.7: Time series plot of the hourly predicted and observed NEE of different validation period for the Taylor Slough (US-Esm) and Lost Creek (US-Los) sites showing impressive validation.

3.3.4 Model sensitivity and uncertainty

Dimensionless sensitivity coefficients of model parameters were analytically derived from Equations (3), (4) and (5) presented in chapter 3 (section 3.2.4) to determine the response of predicted NEE for any changes in parameters for any day and site. Normalization of the sensitivity parameters by the day- and site-specific single reference observations (NEE_{ref}) ensured their applicability for any day and site.

At any diurnal hour, the dimensionless sensitivity coefficient for \hat{a}_0 (the non-oscillatory, principal model component parameter) was 1.0; referring to a constant, linear rate of change in predicted NEE for any change of \hat{a}_0 . The sensitivity coefficients for the oscillatory model

component parameters (\hat{a}_k and \hat{b}_k) varied dynamically with the diurnal hours because of the involvement with sine-cosine functions (Figure 3.8). Among the different diurnal hours at the study wetlands, the sensitivity coefficients of the first harmonic parameters (\hat{a}_1 and \hat{b}_1) varied, respectively, between 0 to 2 and -1.0 to 1, whereas both the second harmonic parameters (\hat{a}_2 and \hat{b}_2) varied between -2.0 to 0 and -1 to 1. The oscillatory component parameters did not show any sensitivity at the reference time since the harmonic function (the dimensionless diurnal curve) was forced to pass through the reference-time observation.

Model uncertainty due to the individual parameter uncertainty was computed by separately changing each parameter value by the corresponding standard deviation (Table 3.2). The Missouri Ozark Forest was considered as an example to demonstrate the computed model uncertainty; since the reference-time observation (NEE_{ref}) can be different on different days, we assumed $NEE_{ref} = 1.0 \mu\text{mol/m}^2/\text{s}$ for simplicity and convenience of demonstrating model responses to the parameter changes for any day. For the US-Esm, changing the absolute temporal mean of \hat{a}_0 (0.053) by the corresponding standard deviation (0.409) resulted in a change of $0.409 \mu\text{mol/m}^2/\text{s}$ in the predicted NEE; changing the harmonic parameters ($\hat{a}_1, \hat{b}_1, \hat{a}_2, \hat{b}_2$) by their respective standard deviations resulted in absolute changes of the predicted NEE by up to 0.822, 0.372, 0.20 and $0.191 \mu\text{mol/m}^2/\text{s}$, respectively. For the US-Los site, changing the parameters ($\hat{a}_0, \hat{a}_1, \hat{b}_1, \hat{a}_2, \hat{b}_2$) by their corresponding standard deviations resulted in absolute changes of the predicted NEE by up to 0.322, 1.07, 0.838, 0.352 and $0.320 \mu\text{mol/m}^2/\text{s}$, respectively. Overall, individual parameter uncertainty was higher in US-Los than US-Esm due to higher standard deviation of the estimated parameters, which was due to fewer observed dispersions of the parameter values at US-Esm.

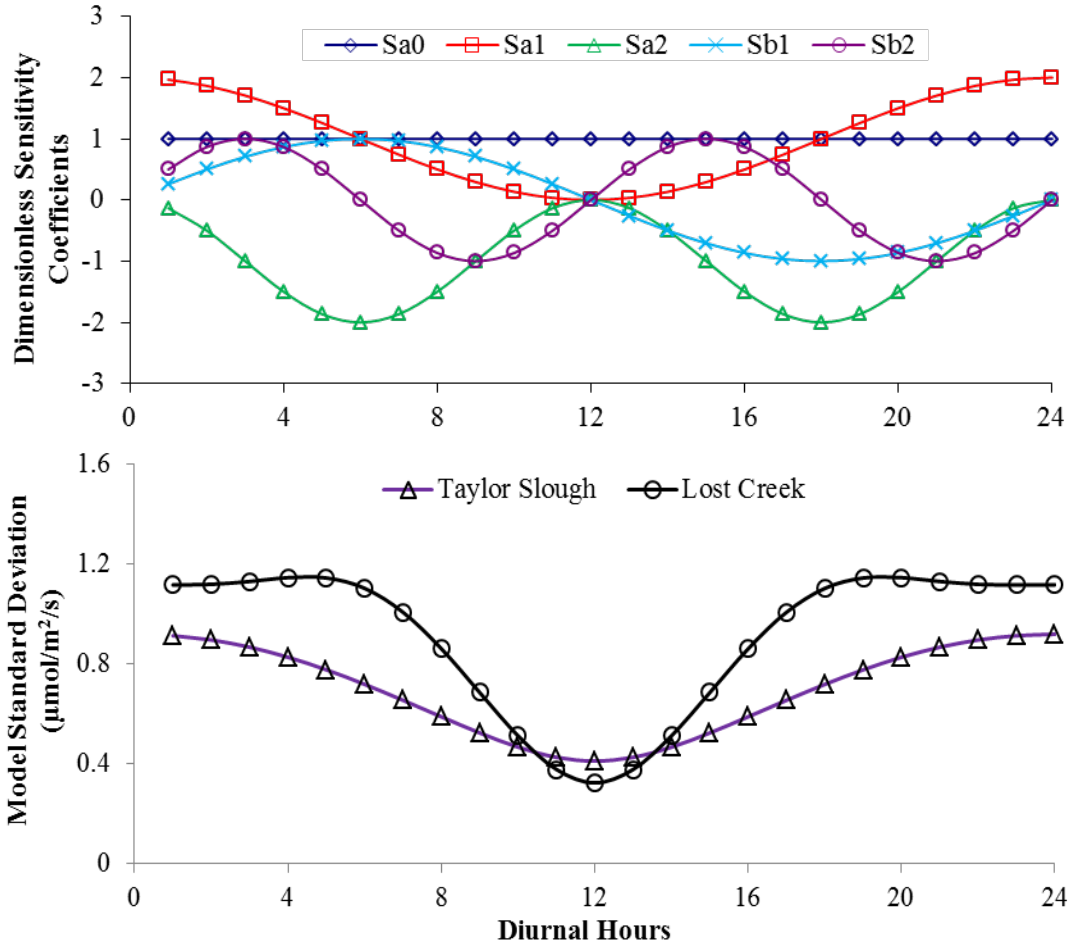


Figure 3.8: Variation of dimensionless parameter sensitivity coefficients ($S_{\hat{a}_0}, S_{\hat{a}_1}, S_{\hat{a}_2}, S_{\hat{b}_1}, S_{\hat{b}_2}$ and $S_{\hat{b}_2}$) with the hours of any diurnal cycle, indicating the model sensitivity for the study wetlands. Since same 12:00 PM was used as a reference time (t_{ref}) for both the wetlands, of dimensionless parameter sensitivity coefficients are identical. ($S_{a0}, S_{a1}, S_{a2}, S_{b1}$, and S_{b2} , respectively, refer to model sensitivity to $\hat{a}_0, \hat{a}_1, \hat{a}_2, \hat{b}_1$ and \hat{b}_2) (upper panel). Bottom panel figure showed standard deviations of the predicted hourly NEE as an overall measure of the model uncertainty for simultaneously changing all parameters by their respective standard deviations at two wetland sites.

However, an overall measure of the model uncertainty for the study wetlands was obtained by computing the standard deviations of the predicted NEE for simultaneously changing all random parameters by their corresponding standard deviations with a combined first order uncertainty analysis (Equation 6, Chapter 2, section 2.2.4) by applying Taylor series expansion

(Mays, 2011) and keeping only the first order terms. For a unit reference observation ($NEE_{ref} = 1.0 \mu\text{mol}/\text{m}^2/\text{s}$), standard deviations of the predicted NEE for different diurnal hours varied between 0.41 and 0.92 $\mu\text{mol}/\text{m}^2/\text{s}$ for US-Esm the standard deviation of predicted NEE varied between 0.32 and 1.15 $\mu\text{mol}/\text{m}^2/\text{s}$ (Figure 3.8).

3.4 Discussion

3.4.1 Success in scaling and potential of finding wetland ecological similitudes

Scaling each diurnal cycle by their corresponding single reference observation brings the collapse of all the different diurnal cycles of wetland NEE into a dimensionless single curve that ultimately leads to a set of parameters that are temporally robust (Figure 3.4); leverages the successful prediction of different NEE cycles of fine resolution. The estimated parameters were insensitive to influential drivers of wetland CO₂ exchanges (PAR, TA and WS) (Figure 3.5) for both the study wetlands (represent coastal freshwater and inland wetland systems) revealing a remarkable attribute of scaling in the domain of one of the most complex systems in nature. Consideration of four years of data while estimating the parameters strengthen the efficiency of the applied scaling technique. Overall, observed success in estimating temporally robust parsimonious set of model parameters and good predictions of different entire diurnal cycles of NEE representing two completely different wetland environment and climate, highlighted the effectiveness of the scaling method to unravel and utilize the underlying ecological similitude (i.e., similarities leading to the parametric reduction) for achieving a useful generalization among different days and feasibly among different wetland types representative of a temporal and geographical gradient in climate, topography and land uses, soil type, hydrology, biogeochemistry and ecology. The wetland ecological simulates, and scaling analysis requires further research for robust estimation and prediction of wetland C budget and sequestration.

3.4.2 Gap-filing of wetland NEE time series and predictions

Proper calibration and validation and the construction of the different diurnal NEE cycles from a corresponding day-specific single reference observation provided the potential of gap-filling in fine resolution (e.g., half-hourly, hourly) time-series of NEE. Presented modeling approach is applicable for the predictions of the NEE diurnal cycles of the relatively ungauged wetlands from available limited observations. The AmeriFlux towers historically encountered approximately 35% missing data annually (includes all types of ecosystems) at the half-hourly scale (Falge et al., 2001), which hinders the precise estimations of CO₂ exchanges and C budget at the longer time-scales (e.g., weekly, monthly, seasonally). The available gap-filling methods such as the empirical mean diurnal variation (MDV) and the semi-empirical non-linear regressions (Falge et al., 2001 for details) are generally difficult to apply in complex wetlands; application of MDV requires corresponding NEE data for 7 to 14 neighboring observations around the missing value for gap filling (Falge et al., 2001) and the semi-empirical regressions can be applied when corresponding data for the relevant mechanistic predictors (e.g., radiation, temperature) are available. So, any of the existing gap-filling methods cannot leverage the limited available data for one time instant to estimate the missing NEE for other times within the same or different diurnal periods. In contrast, presented scaling-based NEE model can provide temporally robust estimations of missing data and predictions of the entire diurnal cycle of NEE based on a corresponding single NEE measurement made at the reference time (mostly noon; can be flexibly chosen between 10:00 AM and 2:00 PM). Once the model is parameterized with continuous NEE data (of preferably multiple days), the scaled NEE model can be utilized to estimate the entire diurnal cycles of NEE for other days by utilizing the corresponding limited observations.

3.4.3 Applicability in different ecosystems and spatiotemporally robust predictions

Each ecosystem has unique climatic, physiological, biogeochemical and ecological characteristics, which impose a formidable challenge to the generalization of the predictive models for different natural systems. Bringing different ecosystems into a comparable modeling framework (time and space transition) for the robust predictions of NEE is an extremely challenging undertaking. Although estimated model parameters in this study are temporally robust, they did not show significant special robustness since estimated parameters of US-Esm were not totally comparable to the parameters of US-Los; presence of high biogeochemical and hydrological gradients (one is coastal freshwater marsh, and other is inland shrub wetland) in the study ecosystems hinder such spatial collapse. However, presented study and the earlier application of this scaling-based ESHA model provided the potential of developing spatiotemporally robust predictive models of NEE. Ishtiaq and Abdul-Aziz (2014) successfully predicted diurnal cycles of NEE for five different deciduous forests using a single set of model parameters derived from the spatiotemporal ensemble of the parameters of five deciduous forests. Further research is, therefore, needed to test the applicability of this scaling-based harmonic (ESHA) model that can offer acceptable predictions of the diurnal NEE cycles for other types of ecosystems apart from deciduous forests and wetlands.

3.4.4 Limitations of model evaluations

The ESHA model parameterization requires at least one (and preferably multiple) well-defined 24-hour diurnal cycle(s) of NEE for the calibration period. Subject to the limited availability of continuous (e.g., hourly) time-series data for wetland NEE, we were able to evaluate the model for only two sites that mainly represented two different freshwater wetland types (coastal vs. inland). Since, US-Los does not have defined diurnal cycle throughout the year (Figure 3.2), only summer growing season (June-September) hourly NEE was considered in this study; however,

US-Esm had relatively defined diurnal cycle for the entire years of simulations. Level-2, gap-filled NEE data of AmeriFlux sites were used to estimate the parameters and calibration of the NEE model. These gaps were filled by the AmeriFlux data community by using the artificial neural network (ANN) and MDV methods (see <http://ameriflux.ornl.gov/>), which could cause input data bias and uncertainty in the ESHA model evaluations. However, consideration of common (two) harmonic structure for different days of the model simulation period should theoretically lessen the potential gap-filling bias. Data bias was further reduced by the use of the temporal ensemble mean (average of daily estimates of parameters) of model parameters during model calibration. Moreover, the successful high-quality validation of the model for the two study sites with the non-gap-filled data proved the applicability of the gap-filled data for the model calibration.

Since ESHA model is completely data-driven, the reliability of the AmeriFlux NEE data is an essential aspect while modeling. A study on AmeriFlux data quality showed the presence of 8.2% relative error for CO₂ flux measurements (Schmidt et al., 2012). At daily scale, AmeriFlux measurement error was around 7% or less during day 12% at night (Berger et al., 2001; Soegaard et al., 2000). However, if there are outliers being present, particularly at the reference time, that might cause inaccurate scaling of the diurnal NEE cycles, which would lead to the unreliable predictions of the hourly NEE. Since the NEE data used in this study passed through the QA/QC checks, it was assumed that the extreme outliers were removed. Further, to minimize the effect of outliers on predictions, the estimated daily parameters of the harmonic model were filtered by retaining values within their respective 1st and 99th percentile ranges and corresponding days were excluded from the calibration dataset.

The temporally ensemble harmonic models could not accurately predict the extreme values of NEE for the study wetlands. (Figure 3.6), which is one of the limitations of temporal

aggregations (averaging) of the harmonic model parameters. However, the day-specific parameters of the harmonic model theoretically can more accurately predict the extreme values of NEE. Furthermore, reliable predictions of high extremes are also unlikely by a conventional process-based model. The extreme values could be better predicted using statistical models specifically designed to predict extremes.

3.5 Conclusions

Presented scaling-based ESHA model is a novel way of utilizing scaling technique in modeling wetland NEE. Successful scaling of the different NEE cycles without any definite trend of the estimated parsimonious set of model parameters lead to the temporally robust predictions of different diurnal cycles of wetland NEE from a corresponding single reference observations. However, use of multiple reference observations instead of single reference observation as the scaling parameter could ideally improve the model predictions, and this assumption is subjected to further research. The model parameters were also nearly insensitive to the individual wetland-specific crucial drivers (PAR, TA, and WS) of ecosystem carbon exchange. The analytically derived parameter sensitivity coefficients and uncertainty measure showed a notable robustness of the model predictions to the standalone and simultaneous perturbations in parameters.

The temporally robust empirical NEE model after parameterization, can be applied for simulating continuous (e.g., hourly) NEE time-series from a single reference observation (or a set of limited observations) without requiring day -specific calibrations. The method can potentially be used for a robust estimation of missing data and gap-filling of observed time-series of NEE or other periodic ecohydrological variables (e.g., latent heat flux, sensitive heat flux) for wetland ecosystems. Since the harmonic model can construct the entire diurnal cycle of NEE, it can lead to an accurate estimate of the longer temporal-scale (e.g., weekly, monthly, seasonally) ecosystem C budget; aiding to derive appropriate strategies of wetland carbon management.

Presented modeling work also highlighted the scope of applying scaling for the investigation of wetland ecological similitude (i.e., similarities leading to the parametric reduction) for achieving a useful generalization among different days and wetlands that lead to the robust estimation of wetland C budget at different spatial and temporal scales.

Acknowledgements

The research was funded by grants from the National Science Foundation CBET Environmental Sustainability (Award No. 1336911), and the National Oceanic and Atmospheric Administration through NERRA (Grant No. NA09NOS4190153). The supports are thankfully acknowledged. We also acknowledge the AmeriFlux sites (US-Esm and UA-Los) for their data records. Water level data for the Lost Creek site were obtained from Dr. Ankur Desai through personal communications. In addition, funding for AmeriFlux data resources was provided by the U.S. Department of Energy's Office of Science.

References

- Abdul-Aziz, O. I., Ishtiaq, K. S., 2014. Robust empirical modeling of dissolved oxygen in small rivers and streams: Scaling by a single reference observation. *Journal of Hydrology*, 511, 648-657.
- Abdul-Aziz, O.I., Wilson, B.N., Gulliver, J.S., 2007. An extended stochastic harmonic analysis (ESHA) algorithm: Application for dissolved oxygen. *Water Resources Research*. 43, doi:10.1029/2006WR005530.
- AmeriFlux site and data exploration system website, 2014 <http://ameriflux.ornl.gov/>
- Aubinet, M., Grelle, A., Ibrom, A., Rannik, Ü., Moncrieff, J., Foken, T., ... Vesala, T., 1999. Estimates of the annual net carbon and water exchange of forests: the EUROFLUX methodology. *Advances in Ecological Research*, 30, 113-175.
- Baldocchi, D. D., 2003. Assessing the eddy covariance technique for evaluating carbon dioxide exchange rates of ecosystems: past, present and future. *Global Change Biology*, 9(4), 479-492.
- Bartlett, K.B., Bartlett, D.S., Harriss, R.C., Sebacher, D.I., 1987. Methane emissions along a salt-marsh salinity gradient. *Biogeochemistry*, 4(3), 183-202.
- Braswell, B. H., Sacks, B., Linder, E., Schimel, D. S., 2005. Estimating ecosystem process parameters by assimilation of eddy flux observations of NEE. *Global Change Biology*, 11, 335-355.
- Bridgman, S. D., Megonigal, J. P., Keller, J. K., Bliss, N. B., Trettin, C., 2006. The carbon balance of North American wetlands. *Wetlands*, 26(4), 889-916.
- Cao, M., Marshall, S., Gregson, K., 1996. Global carbon exchange and methane emissions from natural wetlands: Application of a processes based model. *Journal of Geophysical Research*, 101(D9), 14399–14414.
- Cui, J., Li, C., Sun, G., Trettin, C., 2005. Linkage of MIKE SHE to Wetland-DNDC for carbon budgeting and anaerobic biogeochemistry simulation. *Biogeochemistry*, 72(2), 147-167.
- Dimitrov, D. D., Grant, R. F., Lafleur, P. M., Humphreys, E. R., 2010, Modeling the effects of hydrology on ecosystem respiration at Mer Bleue bog. *Journal Geophysical Research*., 115, G04043, doi:[10.1029/2010JG001312](https://doi.org/10.1029/2010JG001312).
- Dise, N. B., Gorham, E., Verry, E.S., 1993. Environmental Factors Controlling Methane Emissions from Peatlands in Northern Minnesota. *Journal of Geophysical Research*, 98(D6), 10583-10594, doi:10.1029/93JD00160.
- Dyar, T. R., Alhadeff, S. J., 2005. Dissolved oxygen characteristics of Georgia streams. *Proceedings, 2005 Georgia Water Resources Conference*. University of Georgia, Athens, U.S.A., 25–27 April.

- Falge, E., Baldocchi, D.D., Olson, R., et al., 2001. Gap filling strategies for defensible annual sums of net ecosystem exchange. *Agricultural and Forest Meteorology* 107(1), 43-69.
- Fan, Z., David McGuire, A., Turetsky, M. R., Harden, J. W., Michael Waddington, J., Kane, E. S., 2013. The response of soil organic carbon of a rich fen peatland in interior Alaska to projected climate change, *Global Change Biology*, 19, 604–620, doi:10.1111/gcb.12041.
- Frolking, S., Crill, P., 1994. Climate controls on temporal variability of methane flux from a poor fen in southeastern New Hampshire: Measurement and modeling. *Global Biogeochemical Cycles*, 8, 385–397
- Grant, R. F., Desai, A. R., and Sulman, B.N., 2012. Modelling contrasting responses of wetland productivity to changes in water table depth, *Biogeosciences*, 9, 4215-4231, doi:10.5194/bg-9-4215-2012.
- Hargreaves, K. J., Fowler, D., 1998. Quantifying the effects of water table and soil temperature on the emission of methane from peat wetland at the field scale. *Atmospheric Environment*, 32(19), 3275-3282.
- Hollinger, D. Y., Aber, J., Dail, B., et al., 2004. Spatial and temporal variability in forest–atmosphere CO₂ exchange. *Global Change Biology* 10(10), 1689-1706.
- Hondzo, M., Warnaars, T. A., 2008. Coupled effects of small-scale turbulence and phytoplankton biomass in a small stratified lake. *Journal of Environmental Engineering* 134(12), 954-960.
- Hui, D., Wan, S., Su, B., Katul, G., Monson, R., Luo, Y., 2004. Gap-filling missing data in eddy covariance measurements using multiple imputation (MI) for annual estimations. *Agricultural and Forest Meteorology* 121(1): 93-111
- Ishtiaq, K. S., Abdul-Aziz, O. I., 2015. Relative Linkages of Canopy-Level CO₂ Fluxes with the Climatic and Environmental Variables for US Deciduous Forests. *Environmental Management*, 55(4), 943-960.
- Jarvis, A. J., Stauch, V. J., Schulz, K., Young, P. C., 2004. The seasonal temperature dependency of photosynthesis and respiration in two deciduous forests. *Global Change Biology*, 10(6), 939-950.
- Jansson, P.-E., 2012. COUPModel: Model use, calibration and validation, *Transactions of the ASABE*, 55, 1335–1344
- Kotun, K., Renshaw, A., 2013. Taylor Slough Hydrology. *Wetlands*, 1-14.
- Kumar, M. R., Pednekar, S. M., Katsumata, M., Antony, M. K., Kuroda, Y., Unnikrishnan, A. S., 2006. Seasonal variation of the diurnal cycle of rainfall in the eastern equatorial Indian Ocean. *Theoretical and applied climatology*, 85(1-2), 117-122.
- Lund, M., Lafleur, P. M., Roulet, N. T., Lindroth, A., Christensen, T. R., Aurela, M., ... Nilsson, M. B., 2010. Variability in exchange of CO₂ across 12 northern peatland and tundra sites. *Global Change Biology*, 16(9), 2436-2448.

- Lloyd, J., Taylor, J.A., 1994. On the temperature dependence of soil respiration. *Functional Ecology* 8: 315-323
- Metzger, C., Jansson, P. E., Lohila, A., Aurela, M., Eickenscheidt, T., Belelli-Marchesini, L., ... Drösler, M., 2015. CO₂ fluxes and ecosystem dynamics at five European treeless peatlands—merging data and process oriented modeling. *Biogeosciences*, 12(1), 125-146.
- Meyers, S. R., Sageman, B. B., Hinnov, L. A., 2001. Integrated quantitative stratigraphy of the Cenomanian-Turonian Bridge Creek Limestone Member using evolutive harmonic analysis and stratigraphic modeling. *Journal of Sediment Research*, 71(4), 628-644.
- Michaelis, L., Menten, M. L., 1913. Die kinetik der invertinwirkung. *Biochem. z.*, 49(333-369), 352.
- Milne, B. T., Gupta, V. K., Restrepo, C., 2002. A scale invariant coupling of plants, water, energy, and terrain. *Ecoscience*, 9(2), 191-199.
- Moore, T. R., Bubier, J. L., Frohking, S. E., Lafleur, P. M., Roulet, N. T., 2002. Plant biomass and production and CO₂ exchange in an ombrotrophic bog. *Journal of Ecology*, 90(1), 25-36.
- Nestler, J. M., Long, K. S., 1997. Development of hydrological indices to aid cumulative impact analysis of riverine wetlands. *Regulated Rivers: Research & Management*, 13(4), 317-334.
- O'Connor, B. L., Hondzo, M., Dobraca, D., LaPara, T. M., Finlay, J.C., Brezonik, P. L., 2006. Quantity-activity relationship of denitrifying bacteria and environmental scaling in streams of a forested watershed. *Journal of Geophysical Research*, 111, G04014.
- Papale, D., Valentini, R., 2003. A new assessment of European forests carbon exchanges by eddy fluxes and artificial neural network spatialization. *Global Change Biology*, 9(4), 525-535.
- Potter, C. S., 1997. An ecosystem simulation model for methane production and emission from wetlands. *Global Biogeochemical Cycles*, 11(4), 495-506.
- Priestley, M. B., 1981. *Spectral Analysis and Time Series*. Academic Press, London, U.K.
- Reddy, K. R., DeLaune, R. D., 2004. *Biogeochemistry of wetlands: science and applications*. Crc Press.
- Richardson, A. D., Hollinger, D. Y., 2005. Statistical modeling of ecosystem respiration using eddy covariance data: maximum likelihood parameter estimation, and Monte Carlo simulation of model and parameter uncertainty, applied to three simple models. *Agricultural and Forest Meteorology*. 131(3), 191-208.
- Roulet, N., Moore, T., Bubier, J., Lafleur, P., 1992. Northern fens: Methane flux and climate change. *Tellus*, 44B, 100-105.
- Sagerfors, J., Lindroth, A., Grelle, A., Klemedtsson, L., Weslien, P., Nilsson, M., 2008. Annual CO₂ exchange between a nutrient-poor, minerotrophic, boreal mire and the atmosphere. *Journal of Geophysical Research: Biogeosciences* (2005–2012), 113(G1).

- Schedlbauer, J. L., Oberbauer, S. F., Starr, G., Jimenez, K. L., 2010. Seasonal differences in the CO₂ exchange of a short-hydroperiod Florida Everglades marsh. *Agricultural and Forest Meteorology*, 150(7), 994-1006.
- Schedlbauer, J. L., Munyon, J. W., Oberbauer, S. F., Gaiser, E. E., Starr, G., 2012. Controls on ecosystem carbon dioxide exchange in short-and long-hydroperiod Florida Everglades freshwater marshes. *Wetlands*, 32(5), 801-812.
- Sebacher, D.I., Harriss, R.C., Bartlett, K.B., Sebacher, S.M., Grice, S.S., 1986. Atmospheric methane sources: Alaskan tundra bogs, an alpine fen, and a subarctic boreal marsh. *Tellus*, 38B (1), 1-10.
- Sharifi, A., Kalin, L., Hantush, M. M., Isik, S., Jordan, T. E., 2013. Carbon dynamics and export from flooded wetlands: A modeling approach. *Ecological Modelling*, 263, 196-210.
- Smith, K.A., Ball, T., Conen, F., Dobbie, K.E., Massheder, J., Rey, A. 2003. Exchange of greenhouse gases between soil and atmosphere: interactions of soil physical factors and biological processes. *European Journal of Soil Science*, 54(4), 779-791.
- Stauch, V. J., Jarvis, A. J., 2006. A semi-parametric gap-filling model for eddy covariance CO₂ flux time series data. *Global Change Biology*, 12(9), 1707-1716.
- St-Hilaire, F., Wu, J., Roulet, N. T., Frohling, S. E., Lafleur, P. M., Humphreys, E. R., Arora, V., 2010. McGill wetland model: evaluation of a peatland carbon simulator developed for global assessments, *Biogeosciences*, 7, 3517–3530, doi:10.5194/bg-7-3517-2010.
- St-Hilaire, F., Wu, J., Roulet, N. T., Frohling, S., Lafleur, P. M., Humphreys, E. R., Arora, V., 2008. McGill wetland model: evaluation of a peatland carbon simulator developed for global assessments. *Biogeosciences Discussions*, 5(2), 1689-1725.
- Valentini, R., Angelis, P. D., Matteucci, G., Monaco, R., Dore, S., Mucnozza, G. S., 1996. Seasonal net carbon dioxide exchange of a beech forest with the atmosphere. *Global Change Biology*, 2(3), 199-207.
- Van Huissteden, J., Van den Bos, R. M., Marticorena Alvarez, I., 2006. Modeling the effect of water-table management on CO₂ and CH₄ fluxes from peat soils. *Netherlands Journal of Geosciences*, 85, 3–18.
- Wania, R., Ross, I., Prentice, I.C., 2009, Integrating peatlands and permafrost into a dynamic global vegetation model: 1. Evaluation and sensitivity of physical land surface processes, *Global Biogeochemical Cycles*, 23, GB3014, doi:[10.1029/2008GB003412](https://doi.org/10.1029/2008GB003412).
- Walter, B.P., Heimann, M., 2000. A process-based, climate-sensitive model to derive methane emissions from natural wetlands: Application to five wetland sites, sensitivity to model parameters, and climate. *Global Biogeochemical Cycles*, 14, 745– 765.
- Warnaars, T. A., Hondzo, M., Power M.E., 2007. Abiotic controls on periphyton accrual and metabolism in streams: Scaling by dimensionless numbers. *Water Resources Research* 43, W08425.

- Xiao, J., Ollinger, S. V., Frohking, S., Hurtt, G. C., Hollinger, D. Y., Davis, K. J., ... Suyker, A. E., 2014. Data-driven diagnostics of terrestrial carbon dynamics over North America. *Agricultural and Forest Meteorology*, 197, 142-157.
- Xing, Y., Xie, P., Yang, H., Ni, L., Wang, Y., & Rong, K., 2005. Methane and carbon dioxide fluxes from a shallow hypereutrophic subtropical Lake in China. *Atmospheric Environment*, 39(30), 5532-5540.
- Yvon-Durocher, G., Allen, A. P., Bastviken, D., Conrad, R., Gudas, C., St-Pierre, A., ...Del Giorgio, P. A., 2014. Methane fluxes show consistent temperature dependence across microbial to ecosystem scales. *Nature*, 507(7493), 488-491.
- Yurova, A., Wolf, A., Sagerfors, J., Nilsson, M., 2007. Variations in net ecosystem exchange of carbon dioxide in a boreal mire: Modeling mechanisms linked to water table position. *Journal of Geophysical Research: Biogeosciences* (2005–2012), 112(G2).
- Zhang, Y., Li, C., Trettin, C. C., Li, H., Sun, G., 2002. An integrated model of soil, hydrology, and vegetation for carbon dynamics in wetland ecosystems. *Global Biogeochemical Cycles*, 16(4), 9-1.

Chapter 4: Relative linkages of canopy-level CO₂ fluxes with the climatic and environmental variables for US deciduous forests

Abstract

A simple, systematic data-analytics approach was used to determine the relative linkages of different climate and environmental variables with the canopy-level, half-hourly CO₂ fluxes of US deciduous forests. Multivariate pattern recognition techniques of principal component and factor analyses were utilized to classify and group climatic, environmental, and ecological variables based on their similarity as drivers, examining their interrelation patterns at different sites. Explanatory partial least squares regression models were developed to estimate the relative linkages of CO₂ fluxes with the climatic and environmental variables. Three biophysical process components adequately described the system-data variances. The ‘radiation-energy’ component had the strongest linkage with CO₂ fluxes, whereas the ‘aerodynamic’ and ‘temperature-hydrology’ components were low to moderately linked with the carbon fluxes. On average, the ‘radiation-energy’ component showed 5 and 8 times stronger carbon flux linkages than that of the ‘temperature-hydrology’ and ‘aerodynamic’ components, respectively. The similarity of observed patterns among different study sites (representing gradients in climate, canopy heights and soil-formations) indicates that the findings are potentially transferable to other deciduous forests. The similarities also highlight the scope of developing parsimonious data-driven models to predict the potential sequestration of ecosystem carbon under a changing climate and environment. The presented data-analytics provides an objective, empirical foundation to obtain crucial mechanistic insights; complementing process-based model building with a warranted complexity. Model efficiency and accuracy ($R^2 = 0.55$ to 0.81 ; ratio of root-mean-square error to the observed standard deviations, $RSR = 0.44$ to 0.67) reiterate the usefulness of multivariate analytics models for gap-filling of instantaneous data.

4.1 Introduction

Terrestrial ecosystems are the major components of earth's carbon cycle and traditionally regarded as the reservoirs of green carbon. Turbulent land-atmospheric fluxes of vertical CO₂ contribute to ecosystem-scale carbon budget by balancing between the above-ground assimilatory processes (photosynthesis) and both above-ground and below-ground respiratory processes (Heimann and Reichstein 2008; Cao and Woodward 1998; Schimel et al., 2001). The carbon flux-climate feedback process is sensitive to various meteorological, hydrological, and ecological variables at different spatiotemporal scales. Although much research has been conducted to characterize the role of key environmental stressors and climatic variability on land-atmospheric carbon and heat fluxes (e.g., Schmidt et al., 2011), a robust understanding and prediction of turbulent carbon flux dynamics is yet to be achieved (Piao et al., 2013; Morales et al., 2005; Geider et al., 2001). Investigation of the relative effects of climatic and environmental variations on turbulent carbon fluxes is, therefore, an area of active research. In particular, quantification of the relative linkages of climatic and environmental variables with the vertical CO₂ fluxes is an important step towards building relatively simple empirical to complex mechanistic (i.e., process-based) models for robust predictions and management of ecosystem carbon. For example, altering land uses/cover can change surface albedo (light reflectance), net radiation, temperature, soil moisture, heat fluxes, canopy height, and roughness height and friction velocity. Knowledge and insights on the relative carbon flux linkages of climatic and environmental variables would, therefore, help in developing land management strategies and priorities for a maximum uptake (sequestration) and minimum emission of ecosystem carbon.

Numerous process-oriented biosphere models are available for simulation and prediction of carbon fluxes by mechanistically capturing the necessary ecosystem constituents such as plant photosynthesis, respiration and soil biogeochemical processes. Examples of mechanistic models

include IFUSE (Desai 2010), Ecosystem Demography (ED) (Moorcroft et al., 2001), BIOME-BGC (Running and Coughlan 1988; Running and Gower 1991), Equilibrium Boundary Layers (EBL) (Shir and Bornstein 1977), Carbon Tracker (CT) (<http://carbontracker.noaa.gov>), HYBRID (White et al., 1999), ECOSYS (Grant et al., 2012), CENTURY (Glimanov et al., 1997), LINKAGE (Post and Pastor 1996), etc. Most models use photosynthetically active radiation, vegetation index, atmospheric CO₂ concentration, air and soil temperatures, vapor pressure deficit, and soil moisture as inputs (Chen et al., 2011; Sims et al., 2008; Schubert et al., 2012; Li et al., 2007). These models attempt to predict sub-daily to seasonal and inter-annual variability of carbon flux dynamics.

Process-based models are built on numerous scientific hypotheses and their outcome inherently depends on the embedded process formulations and parameterizations (Beer et al., 2010; Keenan 2012). Further, mechanistic modeling generally involves a complex structure; requires data for many input variables that are not always available, and involves a large parameter set making the model predictions quite uncertain. Application of the mechanistic models also requires high computational resources, expert knowledge and specialized skills, hindering their wide-spread applications as tools for a sustainable management of ecosystem carbon under a changing climate and environment.

Data-driven analytic approaches can lay out the foundation for building an appropriate process-based model of warranted complexity. Standard techniques (e.g., linear regressions, non-linear regressions using mainly Arrhenius and Michaelis–Menten equations, artificial neural networks) have already been successfully applied to fill gaps in measured data of ecosystem fluxes and environmental variables (Falge et al., 2001; Carrara et al., 2003; Lloyd and Taylor 1994; Richardson et al., 2006; Hollinger et al., 2004; Barr et al., 2004; Gove and Hollinger 2006; Aubinet et al., 1999; Braswell et al., 2005; Hui et al., 2004; and Stauch and Jarvis 2006). Multi-

scale models were also developed to estimate carbon fluxes from selected ecosystems using remotely sensed and in-situ observations of geographical, environmental and meteorological variables (e.g., Jahan and Gan 2013; Wylie et al., 2007; Makela et al., 2008; Byrne et al., 2005; Oechel et al., 2000). However, lack of knowledge on the general pattern of dominant predictors and their complex interactions, as well as site-specificity of estimated model parameters, still pose as major gaps in data-driven, robust predictions of terrestrial carbon fluxes.

In this paper, I present a simple, systematic data-analytics approach to analyze observational data and determine the relative linkages of different climate and environmental variables with the canopy-level, vertical CO₂ fluxes of eight US deciduous forests. Multivariate pattern recognition techniques such as the principal component analysis (PCA) and factor analysis (FA), in concert with the Pearson correlation analysis, were utilized to classify and group climatic, environmental, and ecological variables based on their similarity as drivers, examining their interrelation patterns and relative influences in different forest sites. Explanatory partial least squares regression (PLSR) models were then developed to estimate the relative linkages of CO₂ fluxes with the climatic and environmental variables. The findings will guide the development of parsimonious empirical models, while informing the building of appropriate mechanistic models, for robust predictions of ecosystem carbon fluxes from the deciduous forests and other similar ecosystems.

4.2 Materials and methods

4.2.1 Study sites

The study sites include the following eight deciduous broad-leaf forests, ranging from the upper mid-western to the northeastern USA (Figure 4.1): (i) Bartlett Experimental Forest, New Hampshire; (ii) Harvard Forest, Massachusetts; (iii) Missouri Ozark, Missouri; (iv) Morgan

Monroe State Forest, Indiana; (v) Ohio Oak Openings, Ohio; (vi) Silas Little Experimental Forest, New Jersey;(vii) University of Michigan Biological Station (UMBS), Michigan; and (viii) Willow Creek Forest, Wisconsin. Although these sites have similarity in vegetation cover, they represent diversity in climate, topography, land uses, soil and hydrologic patterns, etc. Consideration of the relatively large geographical region, therefore, incorporates a potentially high spatial gradient in land-atmospheric carbon flux dynamics.

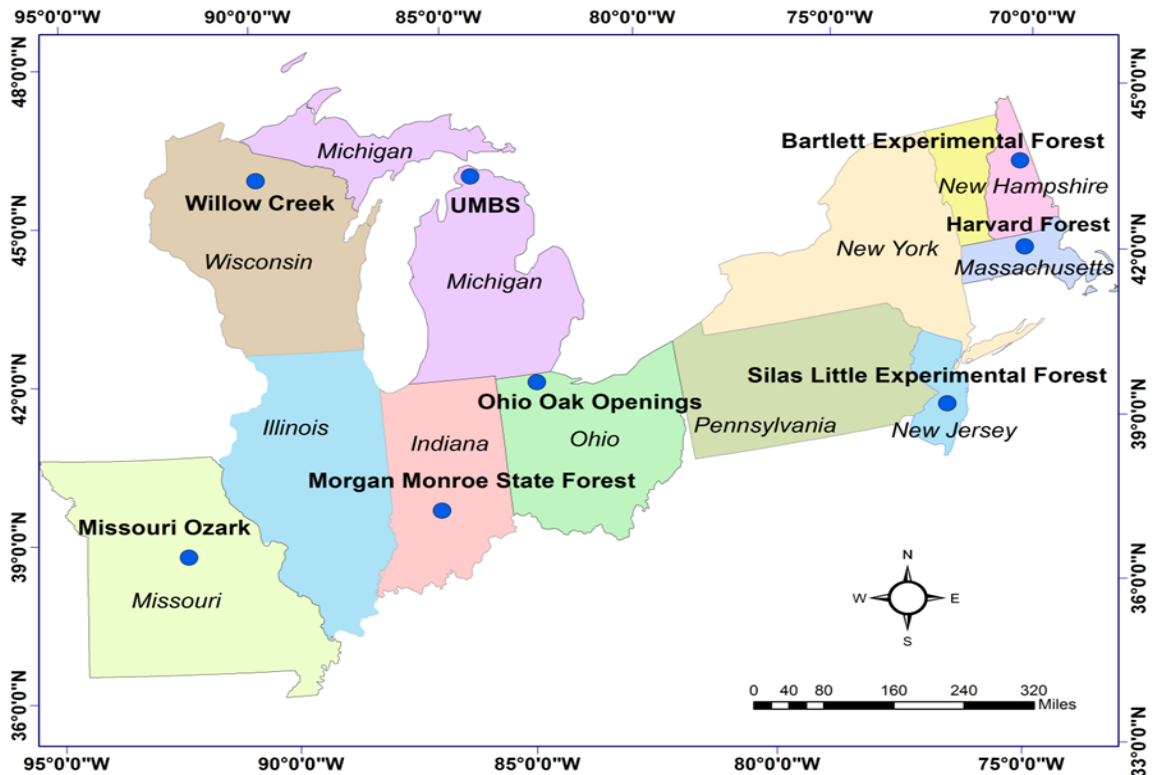


Figure 4.1: Locations of the selected eight deciduous forest study sites

The Bartlett Experimental Forest (canopy height: 19 m) has a rolling to mountainous topography with gentle slope and spodosol soil order; climate is nearly extreme with a minimum daytime temperature of around 90⁰ F during summer and -30⁰ F during winter. Harvard forest (canopy height: 23 m) has a temperate climate and a sandy soil formation in the top layer and sand-gravel-silt at the bottom. Missouri Ozark (canopy height: 24.2 m) and Morgan Monroe State Forests (canopy height: 27 m) share a similar temperate continental climate, while representing

different soil formations (silty loam with a rocky, thin soil cover at Missouri, whereas a combination of clay, loam and limestone residue at the Morgan site). Ohio Oak Forest (canopy height: 24 m) has a humid climate and a flat terrain with sandy mixed soil. Silas Little Experimental Forest (canopy height: 9.52 m) has a cool, temperate climate and a sandy soil formation. The UMBS site (canopy height: 12 m) is characterized by the temperate northern climate and a sandy Entic Haplorthod, glacial till soil. The Willow Creek site (canopy height: 24.2m) represents the northern continental climate with short, moist growing seasons and cold winters, and a soil formation similar to that of the UMBS site. Overall, the selected eight study sites incorporate the potential effects of a variable canopy height (9.52 to 27 m), different climatic regimes (humid to temperate to nearly extreme), and diverse soil morphology on the carbon flux dynamics.

4.2.2 Data sets

Observational data of the half-hourly, canopy-level CO₂ fluxes and the corresponding climate and environmental variables were obtained for different, recent annual cycles (2006-2011) for the study sites from the AmeriFlux network (AmeriFlux website, 2014) (Table 4.2); selection of the 5-year period incorporated the effect of multi-year temporal gradients on carbon flux dynamics. Selected variables were vertical CO₂ flux (measured above the canopy without correcting for underlying storage and advection) (F_{CO_2} ; denoted as F_C in AmeriFlux); ambient CO₂ concentration (C_{CO_2}); net radiation (RN); incoming photosynthetically active radiation (PAR); sensible heat flux (SHF); latent heat flux (LHF) (ecosystem water exchanges); air temperature (TA); soil temperature (TS); vapor pressure deficit (VPD); soil water content (SWC); wind speed (WS); and friction velocity (UST). The TA, WS, PAR, LHF and SHF represent measurement just above the canopy level. The radiation measurements reflect the impacts of cloudiness, which can significantly control plant photosynthesis (Fuentes and Wang, 1999).

Table 4.1: Means and standard deviations (in parenthesis) of the observed carbon fluxes, and climate and environmental variables for the eight broad-leaf deciduous forest sites of USA.

Site	F _{CO2} ($\mu\text{mol}/\text{m}^2/\text{s}$)	UST (m/s)	TA ($^{\circ}\text{C}$)	WS (m/s)	SHF (W/m^2)	LHF (W/m^2)	TS ($^{\circ}\text{C}$)	C _{CO2} ($\mu\text{mol}/\text{mol}$)	VPD (KPa)	SWC (%)	RN (W/m^2)	PAR ($\mu\text{mol}/\text{m}^2/\text{s}$)
Bartlett Forest, NH <i>Year:2009; N = 5597</i>	-3.34 (8.18)	0.44 (0.29)	11.7 (8.90)	1.7 (0.96)	49.00 (112.4)	41.58 (72.38)	9.86 (5.3)	383.4 (17.5)	0.51 (0.51)	35.69 (8.9)	143.4 (224.5)	457 (544.7)
Harvard Forest, MA <i>Year: 2006; N = 4446</i>	-1.37 (7.89)	0.53 (0.30)	8.71 (8.69)	2.19 (0.95)	36.04 (120.48)	30.17 (63.03)	8.61 (6.25)		0.41 (0.39)		95.85 (198.86)	316.1 (481.0)
Missouri Ozark, MO <i>Year:2009; N = 13194</i>	-1.43 (6.71)	0.435 (0.30)	12.97 (9.97)	2.74 (1.29)	33.28 (92.01)	54.61 (92.73)	12.69 (7.31)	390.90 (15.27)	0.635 (0.55)	43.12 (7.91)	107.40 (233.69)	411.1 (599.4)
Morgan Forest, IN <i>Year: 2010; N = 7080</i>	-1.51 (6.63)	0.46 (0.27)	12.29 (11.9)	3.54 (1.28)	32.46 (88.71)	42.86 (90.80)			0.71 (0.65)	35.14 (11.48)	99.45 (221.57)	360.50 (526.7)
Ohio Oak, OH <i>Year: 2009; N = 7729</i>	-4.47 (10.49)	0.49 (0.32)	11.10 (10.6)	2.41 (1.14)	48.48 (106.34)	66.45 (108.8)	11.68 (5.66)	384.1 (16.55)	0.61 (0.59)	6.13 (1.58)	132.6 (236.83)	500.50 (634.5)
Silas Forest, NJ <i>Year:2011; N = 13188</i>	-0.44 (6.07)	0.41 (0.28)	12.69 (10.4)	1.82 (1.08)	36.44 (93.90)	56.29 (109.9)	12.91 (7.41)		0.67 (0.69)		84.43 (194.25)	336.14 (495.9)
UMBS Forest, MI <i>Year:2011; N = 11688</i>	-1.76 (6.97)	0.49 (0.32)	9.85 (11.0)	3.67 (1.65)	38.67 (94.98)	44.10 (89.95)	10.14 (7.10)	389.5 (11.76)	0.49 (0.48)	7.81 (2.41)	113.2 (209.67)	406.6 (539.4)
Willow Creek, WI <i>Year: 2006; N = 8879</i>	-2.21 (6.99)	0.55 (0.29)	7.38 (11.1)	2.52 (1.09)	41.38 (93.70)	40.84 (81.62)	8.87 (7.03)	389.4 (13.00)	0.24 (0.41)	30.86 (4.32)	110.5 (214.21)	398.6 (550.2)

Notes: “Year” and “N” refer to data collection year and sample size, respectively. “Blank” cells in the table indicate the “missing data”. F_{CO2}, UST, TA, WS, SHF, LHF, TS, C_{CO2}, VPD, SWC, RN, and PAR refer, respectively, to CO₂ flux, friction velocity, air temperature, wind speed, sensible heat flux, latent heat flux, soil temperature, ambient CO₂ concentration, vapor pressure deficit, soil water content, net radiation, and photosynthetically active radiation

Choice of the selected study years (2006-2011) was based on the most recent, relatively high availability of reliable data; the set of participatory variables was determined by a preliminary data analysis and leveraging current understanding of carbon flux dynamics in the selected ecosystems. The data sets contain four distinct process partitions of flux, radiation, above ground and below ground environmental variables. Most of the process-based carbon flux prediction models used these components for system representation (Carvalhais et al., 2005; Desai 2010; Chen et al., 2011); some empirical models also adopted a similar approach (e.g., Wylie et al., 2007).

Collected data are classified as level-2, which passed through QA/QC checks. Since data sample for each forest site was very large, gaps in the data matrix were not filled in order to avoid any additional biases in the empirical model building. I applied a two-step procedure to remove the unsuitable data and prepare the final data set for each of the eight sites. First, half-hourly data panels representing gaps for more than two participatory variables were removed from further analysis. This procedure led to the exclusion of half-hourly panel data over a year by around 24% for both Missouri Ozark and Silas Forests, by 33% from the UMBS, by 49% from the Willow Creek, by 55% from the Ohio Oak Forest, by 59% data from the Morgan Forest, by 67% from the Bartlett Forest, and by 74% from the Harvard Forest (see Table 4.1 for the final sample sizes). Although according to Falge et al., (2001), flux towers can encounter around 35% missing data annually at the half-hourly scale, the removal percentages were higher for most stations because some variables (e.g., soil water content), which had not been collected by the eddy covariance method, encountered higher gaps than the flux data. Second, the gap-filtered data for each variable at each station were plotted with time (not shown) to visually check the presence of any unreasonable spikes (i.e., outliers), which were subsequently removed. This secondary filtering led to the removal of 1 half-hourly observation panel for the Silas Forest, 3 observation panels for

the UMBS, 2 observation panels for the Willow Creek, 3 observation panels for the Harvard Forest, and 14 observation panels for the Bartlett Forest. The final data sets (N = 4446 to 13194; see Table 4.1) incorporated the effects of analyzing an equivalent single to multiple seasons on carbon flux dynamics among different stations. Per AmeriFlux sign conventions, positive sign represents upward fluxes (land/forest to atmosphere) of F_{CO_2} , LHF and SHF; negative refers to their downward fluxes (atmosphere to land/forest). In contrast, positive values of radiation (RN and PAR) indicate downward (atmosphere to forest) fluxes and vice versa.

Although photosynthesis and respiration are two different processes and can be dictated by different drivers (e.g., radiation versus temperature), they often share common stressors. Partitioning of F_{CO_2} into photosynthesis and respiration was not considered in this study because the objective was to understand (and quantify) the relative linkages and groupings of different variables influencing the overall diurnal (24-hr) cycle of carbon exchanges. Since the study goal was to determine the relative linkages of mainly climate and environmental variables with the canopy-level vertical carbon fluxes, biological variables such as the canopy leaf-area-index (LAI) were not included in the data matrices. Precipitation was not also included because of the lack of availability of 30-minutes interval; instead SWC was used as a surrogate variable to represent hydrologic effect. I also considered latent and sensible heat fluxes in the data matrix in order to directly incorporate ecohydrologic dynamics.

4.2.3 Data analysis and empirical modeling

I used a simple data-analytics approach that incorporates Pearson correlation analysis, as well as the multivariate PCA, FA and PLSR to emphasize the entire methodology and overall outcomes, rather than the individual analyses or analysis steps. The data-analytics methodology is briefly summarized in a flow diagram (Figure 4.2); the analyses and modeling were done using multiple computation software and programming platforms such as MATLAB, R, and Microsoft Excel.

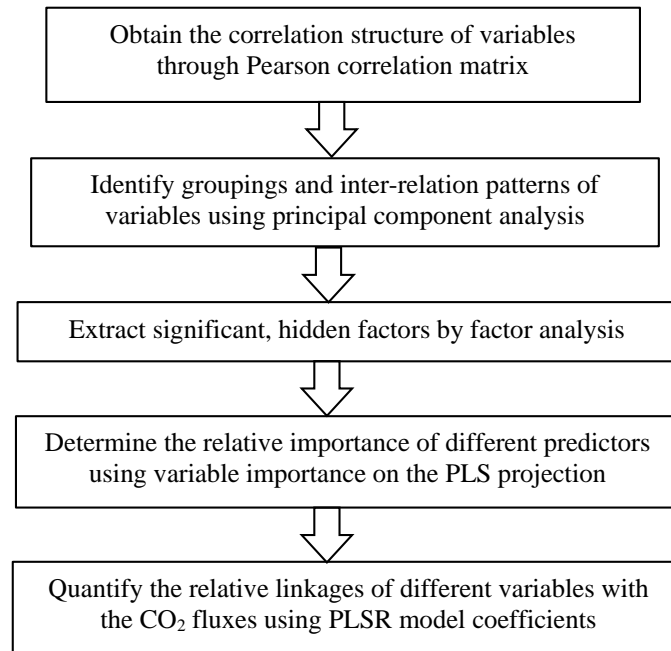


Figure 4.2: The data-analytics methodology to determine the relative carbon flux linkages of different climate and environmental variables

Pearson correlation analysis

Pearson product-moment correlation coefficients were computed to obtain background information on the linear dependency between F_{CO_2} and the climatic/environmental variables for the different deciduous forest sites. Triangular correlation matrices involving all participatory variables were also computed to obtain a preliminary understanding on the multicollinear structure of the data matrix.

Principle component analysis (PCA) and factor analysis (FA)

Multivariate advanced data reduction, interpretation and pattern recognition techniques such as the PCA and FA mine relationships among participating variables in a matrix of independent factors. PCA explains the variance-covariance structure of a data matrix through some orthogonal, linear combinations of original variables, emphasizing data interpretation and reduction in an unsupervised manner (Peres-Neto et al., 2003; Jolliffe 1993; Mahbub et al., 2010). It was previously applied to investigate the spatial representativeness of AmeriFLUX network by

grouping homogeneous areas (Hargrove and Hoffman 2005). A virtue of PCA is that it can unravel relationships hidden in the original data and allows interpretations that are not easy to make using a Pearson correlation matrix. I applied explanatory PCA on the data matrices of the biological, climatic, and environmental variables for the eight study sites. In order to bring different variable units and data sources on a comparable reference scale, data for all the variables were standardized (and made dimensionless to obtain Z-scores) by dividing their instantaneous deviations from the corresponding annual averages by the respective standard deviations (i.e., $Z = \frac{X - \mu}{\sigma}$; where Z = Z-score = normalized variable, X = original variable, μ = annual average of X , and σ = standard deviation of X). First two principal components (PCs) were extracted from the loading matrix (that represents correlation between the PCs and the original variables) and displayed through biplots, which exhibit the possible groupings and interrelations (orientation and correlation strengths) among participatory variables.

FA characterizes the covariance liaisons among many variables with a few rudimental, but unobservable, quantities called factors. FA has been successfully applied for data mining and analysis in many disciplines (Panda et al., 2006; Dragon 2006; and Liu et al., 2003). I applied FA in order to reanalyze the normalized data and verify findings from the PCA by explaining the system variances with fewer latent variables (factors). Individual latent factors were extracted based on an initial eigenvalue criterion (eigen value > 1.0). Additionally, the “varimax” orthogonal rotation was performed, maximizing the sum of the variances of the loading matrices to optimize loading (i.e., correlation) values of the different variables in each factor. Factors extracted thereby were able to describe most of the variances of the data matrices for different study sites.

Partial least squares regression (PLSR)

PLSR is a sophisticated data-driven method to integrate features from a supervised principal component analysis and multiple regressions, explaining the linear relationships between the dependent (i.e., response) variable and independent (i.e., predictor) variables (Wold 1966, 1982). The unique advantage of using PLSR over traditional multiple linear regressions is that it largely eliminates high variability and instability of estimated parameters caused by multicollinearity among predictors. Since the PLSR regression is performed in the transformed orthogonal planes using the independent PLS components by maximally linking data covariance with the response variable, all the predictor variables can be kept in the final model (Kuhn and Johnson 2013). The regression coefficients of the optimal PLSR model are leveraged to compute the regression coefficients (BETA) of the original independent variables by inverting the linear transformations between the PLS components and original variables. Since the issue of multicollinearity is resolved in the PLSR domain, the derived regression coefficients (BETA) of the original variables should ideally be unaffected by any multicollinearity existing in the data matrices.

In order to quantify the relative linkages of climatic and environmental variables with the vertical CO₂ fluxes, normalized (dimensionless) PLSR models were developed by using Z-scores of all participatory variables. The PLSR models were trained (i.e., fitted) and verified (i.e., tested) with half-hourly data (N = 4446 to 13188 among different stations) using the SIMPLS algorithm (de Jong 1993; Hubert and Branden 2003) and a 10-fold cross-validation method (Kuhn and Johnson 2013). The model intercept was “zero” for all sites since Z-score variables were used for the model fitting.

Application of PLSR requires the selection of optimum number of PLS components to ensure minimal prediction error (optimum F statistics) while retaining model stability. The optimal numbers of PLS components were determined using the Akaike Information Criterion

(AIC) (Akaike 1974) and coefficient of determination (R^2), as obtained from the 10-fold cross-validations. Subject to the different sample sizes, a normalized AIC_{FCO_2} was defined to bring all sites to a comparable scale as follows:

$$AIC_{FCO_2}(p) = \ln\left(\frac{SSE_{FCO_2}}{N}\right) + \frac{2p}{N} \quad (1)$$

where p is the total number of the model PLS components, N is the sample size, and SSE_{FCO_2} is the total sum of squared error upon estimation of carbon fluxes (F_{CO_2}). Using the optimum number of PLS components, I employed both the “variable importance in the projection (VIP)” (i.e., PLS-VIP) and “regression coefficients (BETA)” (i.e., PLS-BETA) methods as the complementary approaches to determine the relative importance of different predictors for the model response (Wold et al., 1993, 2001; Chong and Jun 2005). Higher VIP scores and higher regression coefficients indicate more influential predictors in the latent predictor-response matrix; as a rule of thumb, VIP scores exceeding 1.0 can be considered as the most informative predictors for the response (Kuhn and Johnson 2013). Further, the regression coefficients (BETA) of the Z-score PLSR models can represent the relative linkages of the predictor variables (C_{CO_2} , RN, PAR, SHF, LHF, TA, VPD, TS, WS, UST, and SWC) with the response variable (F_{CO_2}).

4.3 Results

4.3.1 Correlation structure of the data

The Pearson correlation coefficients between half-hourly F_{CO_2} and the corresponding climatic/environmental variables were significant at the 95% level of confidence ($\alpha = 0.05$ for a two tailed test) (Table 4.2). Since the participatory variables (e.g., carbon and heat fluxes, temperatures) had both positive and negative values, absolute values of the correlation coefficients were used to describe the linear correspondences of different variables. For all the

study sites, F_{CO_2} demonstrated strong linear correspondences with the biosphere radiations (RN and PAR) ($|r| = 0.61$ to 0.71) and LHF (ecosystem-atmospheric heat and water exchanges due to evapotranspiration) ($|r| = 0.65$ to 0.87); while showing moderate correlations ($|r| = 0.29$ to 0.56) with the SHF (ecosystem-atmospheric heat exchanges due to temperature gradient), as well as with the ambient carbon storage (C_{CO_2}) ($|r| = 0.46$ to 0.57). Moderate linear correspondences of F_{CO_2} were also apparent with the air temperature (TA) ($|r| = 0.28$ to 0.51) and soil temperature (TS) ($|r| = 0.24$ to 0.51) among the different study sites. Vapor pressure deficit (VPD) and available soil moisture (SWC) showed weak to moderate correlations with F_{CO_2} ($|r| = 0.10$ to 0.54 and $|r| = 0.01$ to 0.33 , respectively). In contrast, the linear correspondences of F_{CO_2} with the aerodynamic drivers (WS and UST) were relatively weak ($|r| = 0.03$ to 0.24). Furthermore, the triangular correlation matrices (not shown) for different study sites revealed high mutual correlations among the flux related variables (SHF, LHF, RN and PAR). For example, the correlation coefficient (r) between RN and PAR was 0.99 ; correlation of the radiation variables (PAR and RN) with the SHF and LHF ranged, respectively, from 0.80 to 0.87 and from 0.71 to 0.80 . The temperatures variables (TS, TA and VPD), as well as the velocity variables (WS and UST), were also notably correlated within each group. This indicates the presence of a substantial multicollinearity in the data matrix of the climatic and environmental variables.

4.3.2 Dominant groups and orientation of the variables

For the eight study forests, the first two PCs explained from 61.95% to 75.17% of the total data variances exhibited by the participatory (climatic, environmental, and biological) variables. PCA loading matrices (showing the correlation coefficients between the PCs and the original variables) for different stations are presented through biplots (Figure 4.3 and Figure 4.4).

The first two PCs explained 61.95% of total data variances for the Bartlett Experimental Forest (Figure 4.3a). The orientations and lengths of SHF, PAR, RN, VPD and LHF suggest

Table 4.2: Pearson correlation coefficients between the carbon fluxes (F_{CO_2}) and the climate/environmental variables for the eight deciduous forest study sites (Blank indicates missing data).

Site	RN	PAR	LHF	SHF	C_{CO_2}	TA	TS	VPD	SWC	WS	UST
Bartlett Forest	-0.70	-0.70	-0.72	-0.56	0.46	-0.51	-0.39	-0.10	0.06	0.03	0.03
Harvard Forest	-0.66	-0.67	-0.65	-0.37		-0.39	-0.37	-0.44		-0.03	-0.12
Missouri Ozark	-0.61	-0.61	-0.71	-0.33	0.47	-0.28	-0.24	-0.40	0.17	-0.04	-0.18
Morgan Forest	-0.71	-0.71	-0.84	-0.40		-0.39		-0.38	0.01	0.24	0.06
Ohio Oak	-0.71	-0.70	-0.87	-0.29	0.52	-0.43	-0.33	-0.16	0.26	0.09	0.04
Silas Forest	-0.69	-0.66	-0.77	-0.46		-0.36	-0.26	-0.54		0.05	-0.15
UMBS Forest	-0.67	-0.67	-0.80	-0.44	0.53	-0.38	-0.34	-0.50	0.25	0.07	-0.16
Willow Creek	-0.65	-0.67	-0.84	-0.40	0.57	-0.51	-0.51	-0.42	0.33	-0.20	0.05

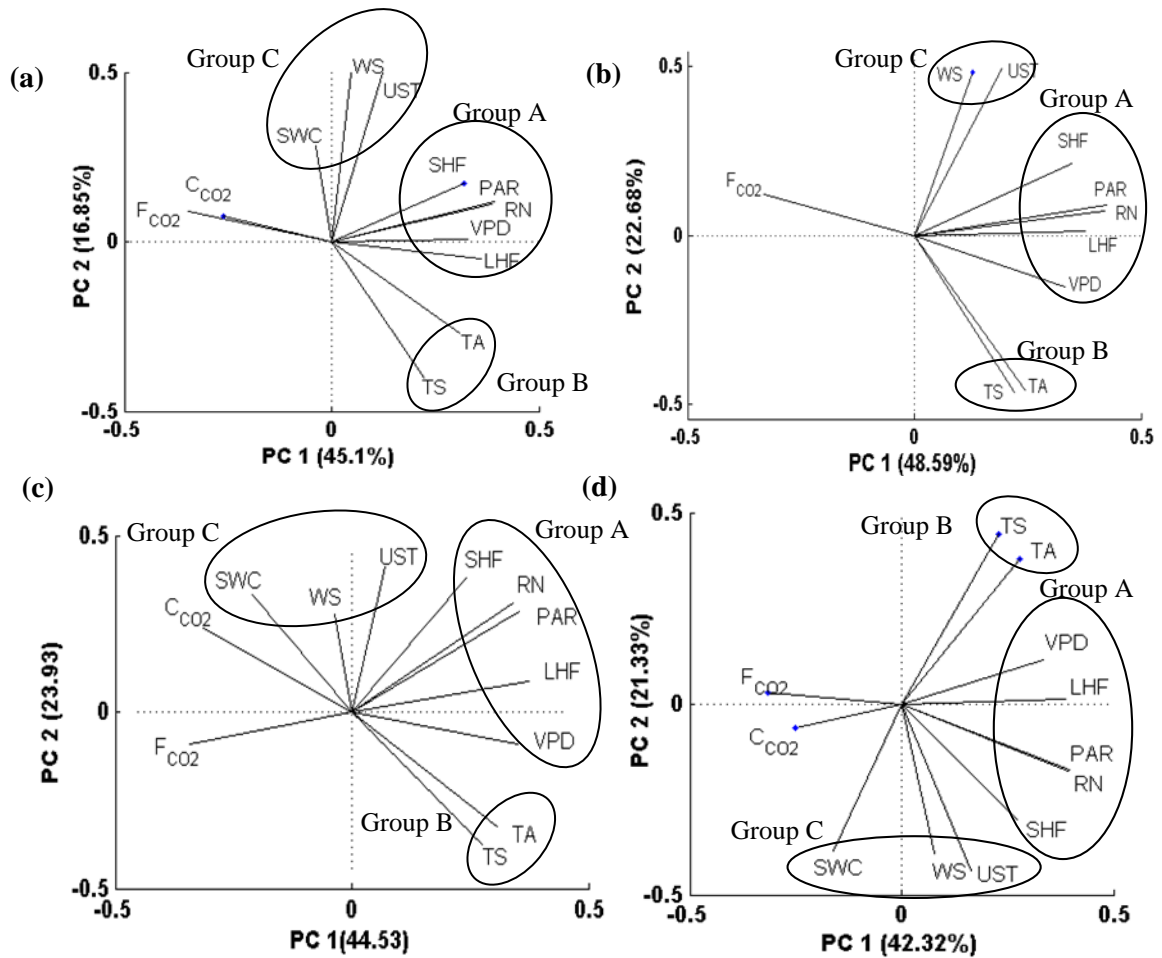


Figure 4.3: Biplots obtained from principal component analysis showing the groupings and interrelation patterns of the climate, environmental, and biological variables for (a) Bartlett Experimental Forest (NH), (b) Harvard Forest (MA), (c) UMBS Forest (MI), and (d) Missouri Ozark Forest (MO). Percent variance explained by each PC is shown in parenthesis.

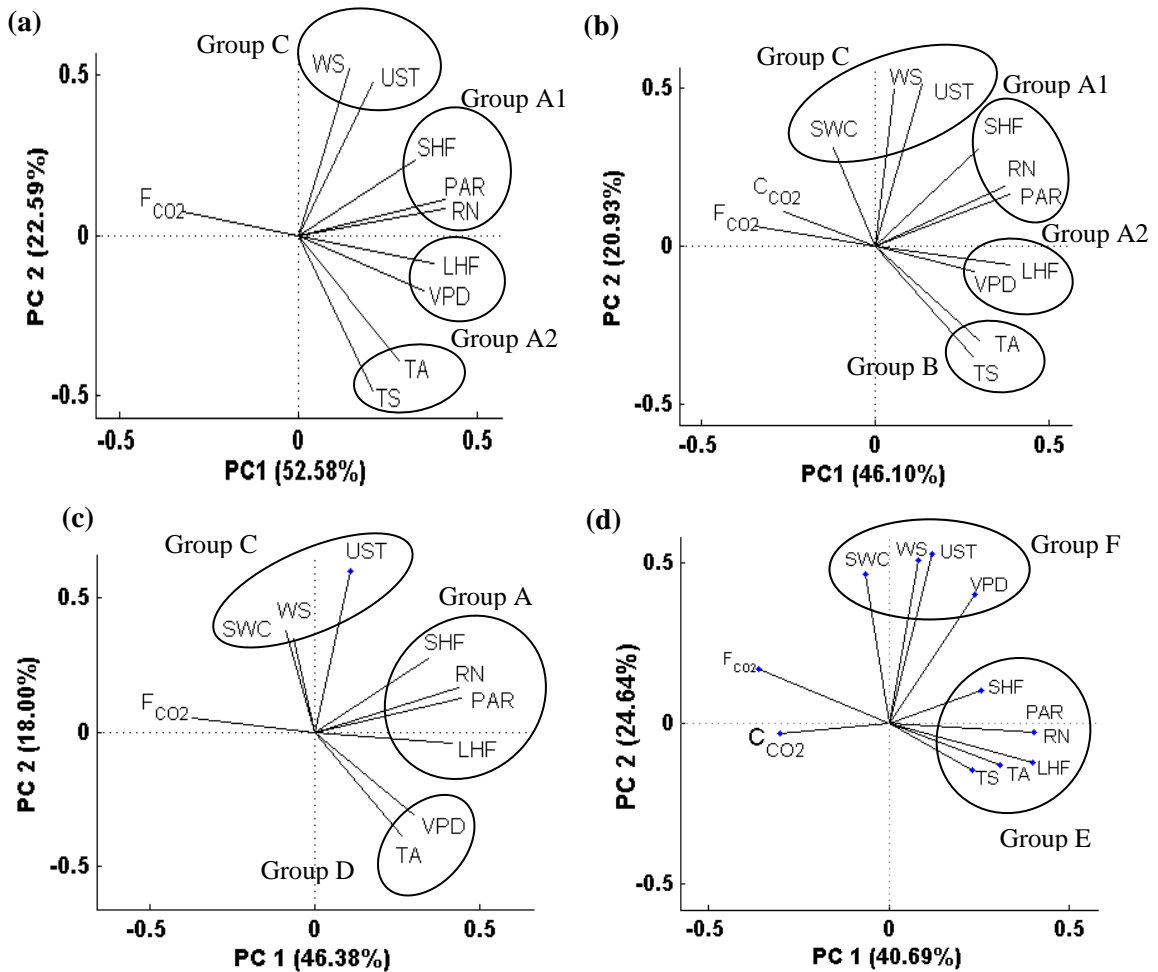


Figure 4.4: Biplots obtained from principal component analysis showing the groupings and interrelation patterns of the climate, environmental, and biological variables for (a) Silas Little Forest (NJ), (b) Willow Creek Forest (WI), (c) Morgan Monroe State Forest (IN), and (d) Ohio Oak Openings Forest (OH). Percent variance explained by each PC is shown in parenthesis.

strong interrelationships, forming a dominant group (A) that highly correlates with F_{CO_2} and C_{CO_2} . TA and TS formed a second group (B), which were moderately correlated with F_{CO_2} . SWC, WS and UST formed the third group (C) that is relatively orthogonal to (i.e., weakly correlated with) F_{CO_2} . Nearly orthogonal orientations of groups A, B and C suggest three different and relatively uncorrelated variance-based clusters of variables hidden in the data; A is dominated by the radiation and flux variables, whereas B and C are dictated by, respectively, the temperature

and aerodynamic variables. The group-A variables and C_{CO_2} , due to their strongly non-orthogonal orientations with F_{CO_2} , are likely the dominant predictors; group-B variables would be the moderately strong predictors, while group-C variables could be the weakest predictors. Similar variable-groupings and interrelation patterns were observed for the Harvard Forest (with missing SWC data) (Figure 4.3b), UMBS Forest (Figure 4.3c), and Missouri Ozark Forest (Figure 5.3d), where the first two PCs explained, respectively, 71.26%, 68.46%, and 63.65% of their total data variances. Although the quadrant-locations of group-B and C were flipped for the Missouri Forest, compared to the other three stations, the relative orthogonality of the three groups, as well as that between F_{CO_2} and group-C variables, were common for all four sites (Figure 4.3).

The Silas Little Forest (with missing SWC data), as well as the Willow Creek Forest, demonstrated a slightly different pattern than that of Bartlett Forest, with the first two PCs explaining, respectively, 75.17% and 67.03% of the total data variances (Figure 4.4). The group-A variables still demonstrated the strongest links with F_{CO_2} ; however, they split into two distinct, non-orthogonal groups of A1 (RN, PAR, SHF) and A2 (VPD, LHF). Similar to the Bartlett Forest, the group-B (temperature) variables showed moderate correspondences, whereas the group-C (wind speed and soil water) variables appeared to be weakly interrelated with the F_{CO_2} . The Morgan Monroe State Forest (the first two PCs explained 64.38% of the total data variance) also showed a slightly different correlation and grouping pattern compared to that of Bartlett or similar forests (Figure 4.4). Although group-C was similar to the previous stations showing a relatively orthogonal orientation with F_{CO_2} , the VPD emerged out of group-A and loaded highly with TA to form a moderately linked group, D; this regrouping could be partly caused by the missing soil temperature data for this station. However, at the Ohio Oak Forest, where the two PCs explained 65.33% of total data variance, all the variables appeared to cluster into two distinct groups of E (SHF, PAR, RN, LHF, TA and TS) and F (SWC, WS, UST and VPD) (Figure 4.4). The group-E has a nearly linear (with 180 degree) orientation with F_{CO_2} , suggesting likely strong linkages

between them; the group-F variables were nearly orthogonally oriented with F_{CO_2} , indicating their possibly weak linkages. The ambient carbon storage (C_{CO_2}) was non-orthogonally oriented with F_{CO_2} , suggesting their strong linkage.

4.3.3 Significant hidden factors

The eigenvalue criterion (eigenvalue >1) led to three independent latent factors for all eight sites (Table 4.3). This means that the extracted three factors adequately summarized the system variance at each site; the first factor explained the most variance (40.69% to 52.6%), then the second factor (16.8 to 24.64 %), while the third factor described the least variance (8.9 to 15.5 %). FA with 'varimax' optimization provided more precise information into the hidden patterns of the data matrices than that shown by the two PCs (see Figure 4.3, 4.4). Since the standardized participatory variables had both positive and negative values, the FA outcomes were interpreted mainly based on the magnitudes (ignoring the positive or negative signs) of the factor loadings.

The radiation and heat fluxes (RN, PAR, SHF, and LHF) generally loaded highly (0.65 to 0.99) on the first factor, which had moderate to high loadings (-0.57 to -0.70) with F_{CO_2} at different stations. The pressure variable (VPD) showed moderate loadings (0.41 to 0.53) on Factor-1 for all but one (Morgan State Forest) sites. The near-canopy carbon storage (C_{CO_2}) showed a moderate loading (-0.38) on Factor-1 only for the Bartlett Forest site. The higher loadings of RN, PAR, SHF and LHF indicate their dictating role on the first hidden factor, which, therefore, can be termed 'radiation-energy' factor. Temperature variables (TA and TS) loaded highly (0.75 to 0.98) with Factor-2, which showed low to moderate loadings (-0.19 to -0.45) with F_{CO_2} . The VPD had moderate to high loadings (0.46 to 0.88) on Factor-2 for all stations; the soil hydrology variable (SWC) loaded moderately (-0.44 to -0.50) on Factor-2 for the Missouri Ozark, Morgan State, and Ohio Oak Forests, while loading slightly (-0.60) for the UMBS Forest. The C_{CO_2} showed

Table 4.3: Dominant latent factors extracted from the data matrices of the eight deciduous forest study sites. “Fac” represents factor

Site	Factors	WS	UST	SWC	VPD	TS	TA	C _{CO2}	SHF	LHF	RN	PAR	F _{CO2}
Bartlett Forest (NH)	Fac 1	0.02	0.20	-0.01	0.48	0.08	0.28	-0.38	0.91	0.65	0.97	0.97	-0.65
	Fac 2	0.04	0.02	-0.14	0.53	0.88	0.94	-0.35	-0.01	0.43	0.20	0.22	-0.38
	Fac 3	0.99	0.89	0.14	0.22	-0.10	-0.02	-0.04	0.06	0.05	0.09	0.12	0.03
Harvard Forest (MA)	Fac 1	0.08	0.25		0.51	0.12	0.16		0.86	0.68	0.95	0.97	-0.63
	Fac 2	-0.06	-0.06		0.49	0.93	0.98		-0.03	0.27	0.19	0.17	-0.29
	Fac 3	0.91	0.97		0.08	-0.07	-0.06		0.20	0.19	0.14	0.16	0.02
UMBS Forest (MI)	Fac 1	-0.07	0.27	-0.05	0.46	0.03	0.14	-0.22	0.85	0.67	0.98	0.98	-0.63
	Fac 2	-0.06	-0.10	-0.60	0.64	0.96	0.96	-0.72	-0.10	0.42	0.12	0.16	-0.34
	Fac 3	0.86	0.96	0.19	0.03	-0.11	-0.07	0.00	0.12	0.07	0.10	0.07	-0.03
Missouri Ozark (MO)	Fac 1	0.03	0.30	-0.07	0.41	0.09	0.17	-0.29	0.83	0.71	0.97	0.97	-0.58
	Fac 2	-0.08	-0.09	-0.50	0.62	0.91	0.97	-0.26	-0.09	0.39	0.18	0.19	-0.20
	Fac 3	0.99	0.78	0.18	0.22	-0.15	0.00	-0.14	0.12	0.08	0.12	0.13	-0.04
Silas Forest (NJ)	Fac 1	0.13	0.29		0.53	0.11	0.23		0.84	0.72	0.95	0.94	-0.67
	Fac 2	-0.09	0.00		0.60	0.91	0.97		0.01	0.37	0.24	0.22	-0.21
	Fac 3	0.99	0.86		0.15	-0.17	0.01		0.22	0.07	0.19	0.23	0.02
Willow Creek (WI)	Fac 1	0.08	0.28	-0.02	0.41	0.18	0.25	-0.34	0.88	0.67	0.95	0.95	-0.57
	Fac 2	-0.10	-0.07	-0.36	0.46	0.95	0.92	-0.39	0.01	0.51	0.21	0.26	-0.45
	Fac 3	0.88	0.96	0.16	0.03	-0.12	-0.08	-0.04	0.21	0.07	0.16	0.14	-0.08
Morgan Forest (IN)	Fac 1	-0.21	0.30	0.04	0.34		0.29		0.82	0.78	0.99	0.98	-0.70
	Fac 2	0.04	-0.14	-0.44	0.88		0.75		0.00	0.32	0.09	0.17	-0.19
	Fac 3	0.74	0.94	0.01	0.04		-0.09		0.12	-0.05	0.03	0.02	0.13
Ohio Oak (OH)	Fac 1	0.07	0.25	-0.03	0.52	0.05	0.25	-0.34	0.83	0.67	0.96	0.97	-0.65
	Fac 2	-0.15	-0.11	-0.49	0.60	0.87	0.90	-0.44	-0.14	0.50	0.21	0.22	-0.37
	Fac 3	0.96	0.89	0.22	0.09	-0.24	-0.07	-0.13	0.25	0.04	0.13	0.11	0.04

moderate to high loadings (-0.44 to -0.72) on Factor-2 for the Ohio Oak and UMBS sites only; of the radiation and heat fluxes, only LHF loaded moderately (0.42 to 0.51) on Factor-2 for the UMBS, Ohio Oak, and Willow Creek Forests. Since temperature related variables and SWC dominated Factor-2, it was termed ‘temperature-hydrology’ factor. Factor-3 (termed ‘aerodynamic’ factor) had a poor loading (0.02 to 0.13) with F_{CO_2} ; only the WS and UST showed notable loadings (0.74 to 0.99) on Factor-3 at different sites.

Similar to the PCA outcomes, the highest loadings of F_{CO_2} with Factor-1 (which explained most of the system variances) refer to the relatively strong linkages of radiation and heat fluxes with the turbulent, vertical CO_2 fluxes; the strong loadings of LHF with both Factor-1 and Factor-2 (ranking second in explaining system variances and F_{CO_2} loadings) indicate a dictating linkage of LHF with F_{CO_2} . Notable double-factor (Factor-1 and 2) associations of VPD and C_{CO_2} also suggest their relatively strong linkages with the vertical carbon fluxes. The very high loadings of TA and TS on Factor-2, as well as that of WS and UST on Factor-3, can also be indicative of their appreciable linkages with F_{CO_2} . The moderate to strong loadings of multiple variables on the ‘radiation-energy’ and ‘temperature-hydrology’ factors reiterate the interrelations (i.e., collinearity) among the radiation, heat fluxes and temperature variables.

4.3.4 Relative carbon flux linkages of the climate and environmental variables

A combination of AIC and R^2 criteria, obtained through a 10-fold cross-validation method, showed that a total of 3 to 5 PLS components led to the optimum PLSR models (minimum AIC_{FCO_2} and maximum R^2), whereas 3 PLS components captured most variations in F_{CO_2} for different sites (Figure 4.5a, b); this is consistent with the FA outcome of 3 independent factors adequately describing the variation in the overall data matrices (Table 4.3). Ranges of optimal model fitting efficiency (R^2 : 0.55 to 0.81) and accuracy (ratio of root-mean-square error to the

standard deviation of observations, RSR: 0.44 to 0.67; mean square error, MSE: 0.19 to 0.45) showed impressive predictions of F_{CO_2} for different study sites (Table 4.4). The model residuals were approximately normally distributed with constant variances (not shown)

The R^2 indicated the amount of observed data variance explained by the model (i.e., model efficiency), whereas the RSR (see notes of Table 4.4 for the mathematical expression) and MSE indicated the accuracy of model fitting. Moriasi et al. (2007) provided a range of RSR

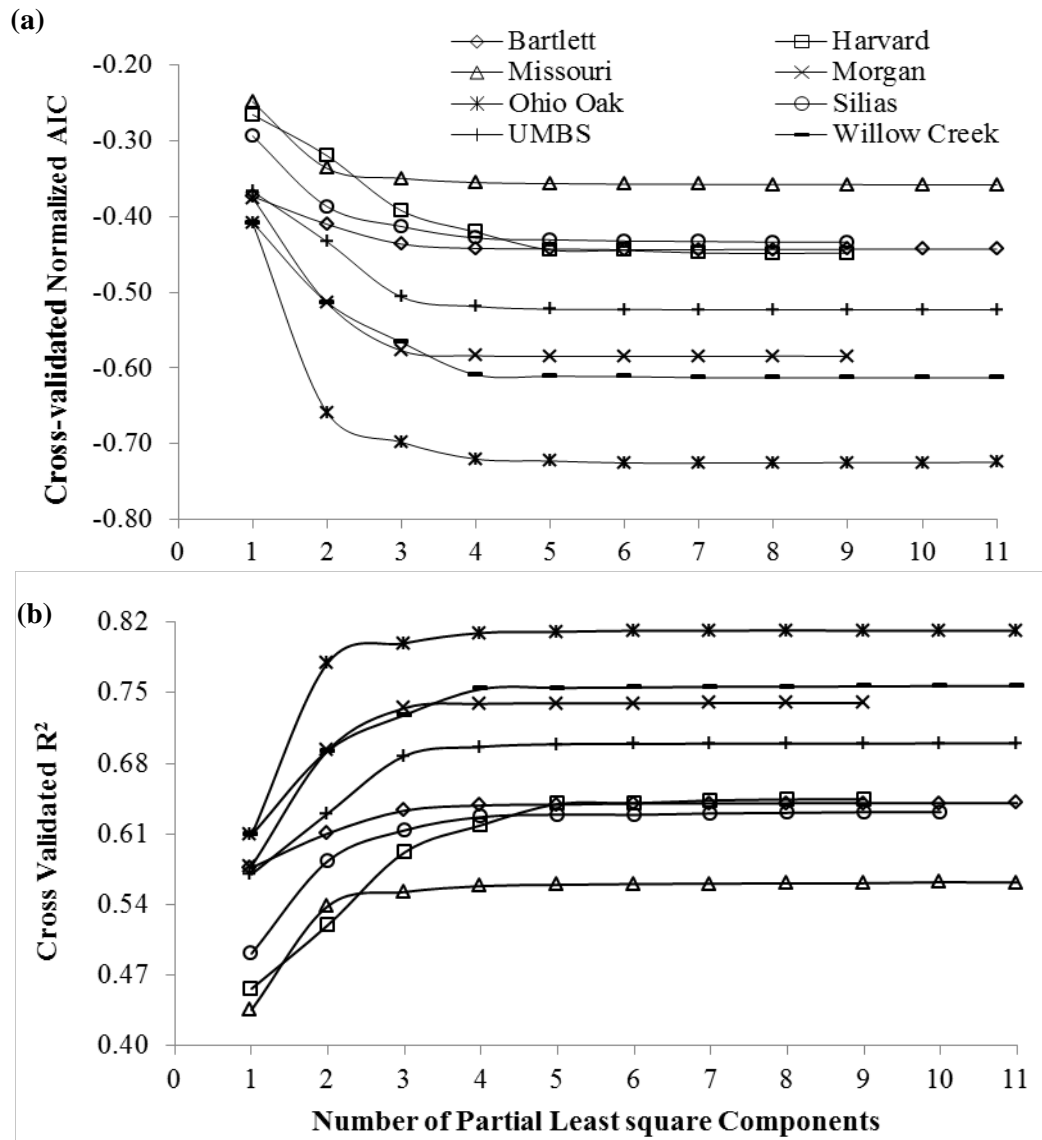


Figure 4.5: Plot of cross-validated (a) normalized AIC and (b) fitting efficiency (R^2) for F_{CO_2} with the number of incorporated partial least squares (PLS) components

values for the evaluation of model accuracy; an RSR from 0 to 0.50 indicates a perfect to very good model, from 0.5 to 0.6 indicates a good model, and from 0.6 to 0.7 refers to a satisfactory model; a model with $RSR > 0.70$ is considered unsatisfactory.

The variable importance in the PLS projection (VIP) scores and regression coefficients (BETA), as obtained with the optimum number of PLS components, quantified the relative linkage of each predictor with the model response (F_{CO_2}) (Figure 4.6; Table 4.4). Since the Z-scores of the participatory variables had both positive and negative values, interpretation of the type (e.g., mutual increase or decrease) of their relative carbon flux linkages based on the positive or negative sign of BETA (β) would be potentially misleading; the predictive influence of individual predictors of F_{CO_2} was, therefore, evaluated by comparing the absolute values of the associated β . Total relative linkages of the ‘radiation-energy’ (β_{RHF}), ‘temperature-hydrology’ (β_T), and ‘aerodynamic’ (β_W) components were computed, respectively, as:

$$\beta_{RHF} = \sqrt{\beta_{RN}^2 + \beta_{PAR}^2 + \beta_{LHF}^2 + \beta_{SHF}^2}; \beta_T = \sqrt{\beta_{TA}^2 + \beta_{TS}^2 + \beta_{VPD}^2 + \beta_{SWC}^2}; \text{ and}$$

$$\beta_W = \sqrt{\beta_{WS}^2 + \beta_{UST}^2}.$$

Bartlett Experimental Forest

The LHF, RN, PAR, SHF, VPD had VIP scores higher than unity (1.0) and were potentially strongly linked with F_{CO_2} (Figure 4.6a); smaller VIPs of TA, C_{CO_2} , and TS indicated their likely moderate linkages with F_{CO_2} ; much lower VIP scores indicated relatively weak linkages of F_{CO_2} with UST, WS and SWC. Model coefficients (β) revealed almost similar linkage pattern with some exceptions (Table 4.4); F_{CO_2} showed relatively strong linkage with C_{CO_2} and relatively weak linkages with SHF and TA. LHF was the strongest predictor in regression, exhibiting around 2 times stronger linkages with F_{CO_2} than that of RN, PAR and C_{CO_2} ; 2.5 to 3.5 times stronger linkages than that of VPD and TS; and around 9 to 36 times stronger linkages than that of the

aerodynamic and soil moisture variables (WS, UST, and SWC). Based on the ratio of β_{RHF}/β_T and β_{RHF}/β_W , the ‘radiation-energy’ component had around 2.5 and 11 times stronger linkages with the carbon fluxes than that of, respectively, the ‘temperature-hydrology’ and ‘aerodynamic’ components.

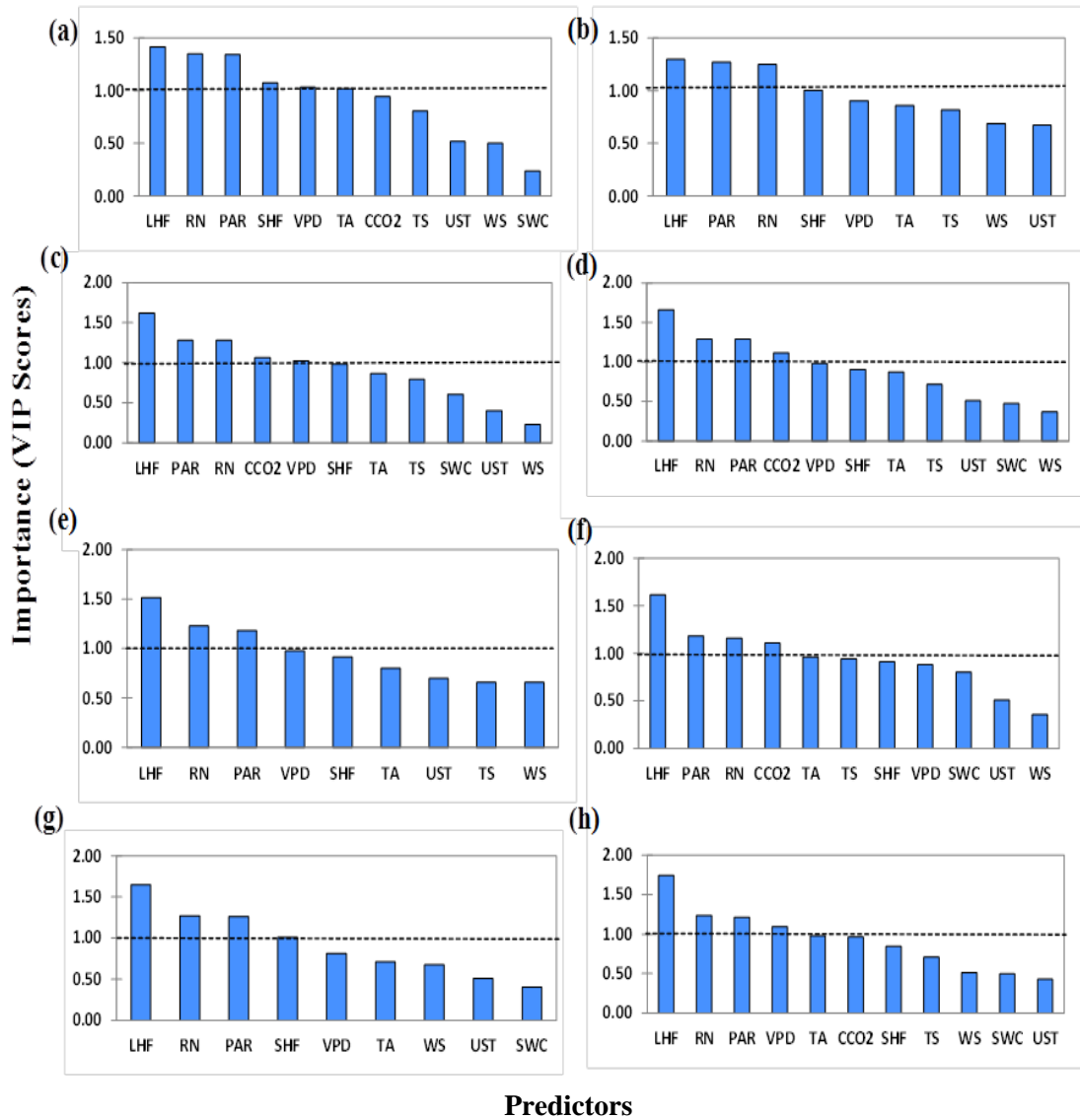


Figure 4.6: Variable importance on the partial least squares projection (VIP) of different predictors for the response variable (F_{CO_2}) for (a) Bartlett Experimental Forest (NH), (b) Harvard Forest (MA), (c) UMBS Forest (MI), (d) Missouri Ozark Forest (MO), (e) Silas Little Forest (NJ), (f) Willow Creek Forest (WI), (g) Morgan Monroe State Forest (IN), and (h) Ohio Oak Openings Forest (OH). Dashed line indicates predictors having VIP score greater than unity (1.0).

Table 4.4: Coefficients (β) of the normalized (dimensionless) PLSR models of carbon fluxes (F_{CO_2}) for different deciduous forest study sites.

Predictor Variables	Bartlett Forest (NH)	Harvard Forest (MA)	UMBS Forest (MI)	Missouri Ozark (MO)	Silas Forest (NJ)	Willow Creek (WI)	Morgan Forest (IN)	Ohio Oak (OH)
RN	-0.20	-0.53	-0.19	-0.17	-0.22	-0.19	-0.23	-0.26
PAR	-0.17	-0.60	-0.15	-0.16	-0.15	-0.14	-0.18	-0.20
LHF	-0.36	-0.30	-0.54	-0.44	-0.58	-0.69	-0.64	-0.67
SHF	-0.08	0.68	0.07	0.11	0.11	0.22	0.24	0.23
VPD	0.14	0.11	0.09	0.04	-0.10	0.16	0.09	0.17
C_{CO_2}	0.18		0.26	0.28		0.15		0.13
TA	-0.05	0.01	0.07	0.06	0.07	0.00	-0.06	0.07
TS	-0.10	-0.06	0.03	0.02	0.10	0.00		0.00
WS	0.04	-0.06	0.11	0.07	0.06	0.02	0.04	0.10
UST	0.01	0.15	0.01	0.05	0.09	-0.04	0.02	-0.04
SWC	-0.03		-0.01	-0.03		0.05	-0.07	0.01
<i>PLS components</i>	3	5	3	3	4	4	3	4
R^2	0.64	0.64	0.69	0.55	0.63	0.75	0.74	0.81
RSR	0.60	0.59	0.56	0.67	0.61	0.50	0.51	0.44
MSE	0.37	0.35	0.31	0.45	0.37	0.25	0.26	0.19
β_{RHF} / β_T	2.5	8.3	5.0	6.3	4.1	4.5	5.8	4.2
β_{RHF} / β_W	11	6.8	5.4	5.9	6.0	18.8	16.6	7.2

Notes: Blank indicates missing data; RSR, the ratio of root-mean-square error to the standard deviation of observations, was calculated as: $RSR = \sqrt{\sum_{i=1}^N (F_{CO_2,i,mod} - F_{CO_2,i,obs})^2 / N} / \sigma_{F_{CO_2,obs}}$, where N is the total number of standardized observations of F_{CO_2} , $F_{CO_2,i,obs}$ and $F_{CO_2,i,mod}$ are the i^{th} observed and predicted value of standardized F_{CO_2} (respectively), and $\sigma_{F_{CO_2,obs}} = 1.0$ is the standard deviation of the observed, normalized F_{CO_2} ; and MSE is the mean square error of predicted, normalized F_{CO_2} . The total relative linkages of the ‘radiation-energy’ (β_{RHF}), ‘temperature-hydrology’ (β_T), and ‘aerodynamic’ (β_W) components were computed, respectively, as: $\beta_{RHF} = \sqrt{\beta_{RN}^2 + \beta_{PAR}^2 + \beta_{LHF}^2 + \beta_{SHF}^2}$; $\beta_T = \sqrt{\beta_{TA}^2 + \beta_{TS}^2 + \beta_{VPD}^2 + \beta_{SWC}^2}$; and $\beta_W = \sqrt{\beta_{WS}^2 + \beta_{UST}^2}$.

Harvard Forest

The VIP scores and PLSR coefficients slightly differed for this site. The radiation and heat flux variables of PAR, RN, LHF, and SHF had the higher VIP scores, referring to their strong linkages with F_{CO_2} (Figure 4.6b). Lower VIP scores of the temperature and aerodynamic variables (VPD, TA, TS, WS, and UST) indicated their relatively weak to moderate linkages with F_{CO_2} . In terms

of regression coefficients (β ; Table 4.4), F_{CO_2} had relatively high linkages with SHF, RN, PAR, and LHF; moderate linkages with VPD and UST; and low linkages with TA, TS, and WS. Unlike other study sites, SHF appeared to be the strongest predictor of F_{CO_2} , although VIP scores indicated stronger dominance of LHF over SHF. Overall, the ‘radiation-energy’ group had approximately 7 and 8 times stronger linkages with F_{CO_2} than that of the ‘aerodynamic’ and ‘temperature-hydrology’ groups, respectively. The relatively high coefficient of UST, compared to other sites, indicated a notable influence of canopy layer turbulent mixing on carbon flux transfer at this site.

UMBS

The VIP scores of LHF, PAR, RN, C_{CO_2} , VPD, and SHF were greater than (or equal to) 1.0, suggesting their strong relative linkages with F_{CO_2} ; lower VIP scores indicated relatively weak to moderate linkages of F_{CO_2} with TA, TS, SWC, UST, and WS (Figure 4.6c). Relative influence (β) of the PLSR model variables were mostly similar to that of VIPs (Table 4.4). LHF was the strongest predictor of F_{CO_2} ; around 2 to 3.5 times stronger than that of the C_{CO_2} , RN and PAR. In contrast, the variables of WS, VPD, TA and SHF showed around 5 to 7.5 times lower linkages with F_{CO_2} . Based on β , F_{CO_2} was very weakly linked with TS, UST and SWC. The ‘radiation-energy’ component had around 5 times stronger linkages with F_{CO_2} than that of the ‘temperature-hydrology’ and ‘aerodynamic’ components.

Missouri Ozark Forest

The higher VIP scores of LHF, RN, PAR, C_{CO_2} , and VPD indicate that these variables were potentially strongly linked with F_{CO_2} ; lower VIPs of the other variables refer to their low to moderate linkages with the carbon fluxes (Figure 4.6d). These results were consistent with the regression modeling outcomes with the exceptions of SHF and VPD, which showed, respectively, relatively high and low coefficients with F_{CO_2} . The LHF was the strongest predictor of carbon fluxes, showing around 1.5 times stronger linkages than that of C_{CO_2} ; 2.5 to 4 times stronger

linkages than the PAR, RN and SHF; and 6 to 8 times stronger linkages than TA, WS, and UST. The SWC and TS did not show noteworthy linkages with F_{CO_2} . Relatively lower model fitting efficiency and accuracy ($R^2 = 0.55$; $RSR = 0.67$; $MSE = 0.45$) also indicate the presence of more complicated (exceedingly nonlinear) carbon processes. Overall, the ‘radiation-energy’ group had approximately 6 times stronger linkages with F_{CO_2} than that of the ‘aerodynamic’ or ‘temperature-hydrology’ group (Table 4.4).

Silas Little Experimental Forest

Based on the common outcomes of VIP scores (Figure 4.6e) and PLSR coefficients (Table 4.4), the LHF, RN and PAR had high linkages with F_{CO_2} , while the VPD, SHF, TA, TS, WS, and UST were relatively low to moderate carbon flux linkages. Per β , LHF was the strongest carbon flux predictor; around 2.5 to 4 times stronger than the RN and PAR; 5 to 6 times stronger than SHF, VPD, TS, and UST; and 8 to 10 times stronger than TA and WS. The ratio of β_{RHF}/β_T and β_{RHF}/β_W showed that the ‘radiation-energy’ component had around 4 and 6 times stronger linkages with F_{CO_2} than that of, respectively, the ‘temperature-hydrology’ and ‘aerodynamic’ components.

Willow Creek Forest

Both the VIP scores (Figure 4.6f) and regression coefficients (Table 4.4) showed stronger linkages of carbon fluxes with the LHF, PAR, RN, C_{CO_2} , SHF, and VPD for this site; the soil moisture (SWC), wind speed (WS) and friction velocity (UST) had relatively low carbon flux linkages. The air and soil temperatures (TS and TS) showed negligible linkages with F_{CO_2} . Based on β , the predictive influence of LHF was around 3 to 4 times stronger than that of RN, PAR, SHF, C_{CO_2} ; and 13 to 17 times stronger than SWC and UST. Overall, the ‘radiation-energy’ group had approximately 4.5 and 19 times stronger linkages with F_{CO_2} than that of the ‘temperature-hydrology’ and ‘aerodynamic’ group (Table 4.4).

Morgan Monroe State Forest

The radiation and heat flux variables of LHF, RN, PAR and SHF had high VIP scores (>1.0), indicating their potentially strong linkages with the carbon fluxes; small VIP scores indicated relatively weak linkages of F_{CO_2} with VPD, TA, WS, UST, and SWC. The PLSR model coefficients (β) (Table 4.4), as well as the outcomes of PCA (Figure 4.4c) and FA (Table 4.3), revealed nearly an identical linkage pattern. The LHF was the strongest predictor in regression, exhibiting around 2.5 to 3.5 times stronger linkages with F_{CO_2} than that of RN, PAR and SHF; 7 to 10 times stronger linkages than that of VPD, SWC, and TA; and around 16 times stronger linkages than WS. Based on β_{RHF}/β_T and β_{RHF}/β_W , the ‘radiation-energy’ component had around 6 and 16.5 times stronger linkages with the carbon fluxes than that of, respectively, the ‘temperature-hydrology’ and ‘aerodynamic’ components.

Ohio Oak Openings

The PLSR model had the highest fitting accuracy (MSE = 0.19; RSR = 0.44) and efficiency ($R^2 = 0.81$) for this site. Both the VIP scores (VIP >1 ; Figure 4.6h) and regression coefficients (Table 4.4) showed relatively strong linkages of carbon fluxes with the LHF, PAR, RN, and VPD; the ambient carbon concentration (C_{CO_2}) had a moderate linkage, whereas the soil moisture (SWC) and friction velocity (UST) had very low carbon flux linkages. Although the sensible heat flux (SHF) and air temperature (TA) had moderate VIP scores, the associated PLSR model coefficients (β), as well as the PCA biplot (Figure 4.4d) and FA results (Table 4.3), showed relatively strong influence of SHF and weak influence of TA in predicting F_{CO_2} . However, contrary to the VIP scores and the PCA and FA results, the wind speed (WS) had a moderately high regression coefficient (i.e., moderate carbon flux linkage), while the soil temperature (TS) showed a negligible coefficient (little linkage) with F_{CO_2} . Based on β , the predictive influence of LHF was around 2.5 to 4 times stronger than that of the RN, PAR, SHF and VPD; 5 to 6.5

times stronger than C_{CO_2} and WS; and 9.5 to 17 times stronger than that of TA and UST. Overall, the ‘radiation-energy’ group had approximately 4 and 7 times stronger carbon flux linkages than that of the ‘temperature-hydrology’ and ‘aerodynamic’ group, respectively (Table 4.4).

4.4 Discussion

4.4.1 Linking vertical carbon fluxes with the climatic and environmental variables

For all eight deciduous forest sites, the ‘radiation-energy’ component of RN, PAR, SHF and LHF was strongly linked with the canopy-level CO_2 fluxes. Previous research (e.g., Jung et al., 2011; Schmidt et al., 2011; Morales et al., 2005; Zhang et al., 2005; Baker et al., 2003; Sellers et al., 1997) also reported a similar finding. However, most studies considered the latent heat flux (LHF) and sensible heat flux (SHF), along with F_{CO_2} , as the response variables as functions of some common drivers such as the radiation, temperature, vapor pressure, etc. In contrast, unlike Melesse and Hanley (2005), this study included the two heat fluxes in the matrix of predictor variables; quantifying their relatively high linkages with F_{CO_2} exchanges within a large set of climatic and environmental variables.

The mutual correlations among the three fluxes of F_{CO_2} , SHF and LHF could partly stem from the common dynamic term (vertical wind speed fluctuations) of their eddy covariance measurement equations (Launiainen et al., 2005); however, their interrelationships identified in this study were mostly process-oriented. The carbon and energy fluxes together represent the ecosystem’s biological exchanges with atmosphere (Baldocchi et al., 2001). Heat fluxes (LHF and SHF) help to maintain balance in atmospheric radiation through evapotranspiration (ET) and turbulent energy diffusion (Sellers et al., 1997). Plants’ stomata tend to close with increasing transpiration and LHF slows down to maintain ecosystem water budget, indirectly affecting carbon flux exchanges (Heber et al., 1986). Furthermore, the relative weight (i.e., ratio) between

the canopy-level photosynthesis and ET is a moderate function of atmospheric humidity deficit (Baldocchi and Meyers 1998), indicating ET-control on high rate of canopy growth.

Although the ambient CO₂ concentration (C_{CO2}) was not available for three of the eight selected study sites, it showed relatively moderate to strong linkages with carbon fluxes (F_{CO2}) for all available sites. Elevated C_{CO2} stimulates the photosynthesis by increasing the carboxylation and oxygenation, leading to a fast plant growth and ultimately increasing the litter production and soil carbon storage (Masle 2000). Furthermore, high atmospheric carbon concentrations often facilitate more efficient use of available soil water for plant growth and productivity (Schlesinger 1999). In contrast, Drake et al., (1999) reported a negative functional relation between plant respiration and C_{CO2}.

The ‘temperature-hydrology’ component (formed by TA, TS, VPD and SWC) had a moderate linkage with the vertical carbon fluxes for almost all study forests. Water availability, temperature and light can play a significant role in plants’ energy and water exchanges by transforming stomatal aperture (short-term) and density (long-term) (Haworth et al., 2011). High vapor pressure deficit (VPD) can lead to stomatal closure (Loescher et al., 2003), affecting the plant-atmospheric energy (latent heat) and carbon fluxes. VPD also loaded appreciably in the ‘radiation-energy’ factor (Table 4.3), which is consistent with previous studies (e.g., Lund et al., 2010). Although soil moisture (SWC) at most study sites grouped with the velocity variables in the PCA biplots (Figure 4.2, 4.3), more detailed (3-dimensional) information emerged in FA analysis showing its appropriate association with the ‘temperature-hydrology’ component (Table 4.3). The SWC can contribute to the plants’ photosynthesis by influencing the water potential difference between the tree leaves and root system. However, the five layer analysis (Pearson correlation, PCA, FA, PLSR-VIP, and PLSR-BETA) with half-hourly data of different years from different US deciduous forests indicate a relatively low predictive influence of soil moisture on

vertical CO₂ fluxes for the small time-scale. The ‘aerodynamic’ component (WS, UST) was also relatively weakly linked with the carbon flux exchanges. Mechanistically UST and WS are not direct contributors of carbon fluxes (Wilson et al., 2002); rather they influence the boundary layer vertical mixing to facilitate transport. Future studies should focus into the attributions of these low carbon flux linkages of SWC, WS, and UST; for example, whether aggregating data in larger time-scales (e.g., daily, weekly, monthly, yearly) reveal a more notable linkage of SWC with the F_{CO₂}.

Despite the gradients of a variable canopy height (9.52 to 27 m), different climatic regimes (humid to temperate to nearly extreme), and diverse soil morphology among the study sites, the relative linkages of major process components, as well as individual climate and environmental variables, with F_{CO₂} were essentially similar. The findings of the relative carbon flux linkages, using a simple method on half-hourly data for exclusively deciduous forests, complement the findings of Schmidt et al., (2011) that applied a complex neural network approach to analyze daily averaged data for different type of ecosystems, including five deciduous forests. For example, both studies conclude that air temperature had a much lower effect on turbulent carbon fluxes than that of other variables such as the radiations; the same is true for the relatively weak carbon flux linkage of wind and friction velocities. However, Schmidt et al., (2011) reported notably strong linkages of daily precipitation and soil temperature with F_{CO₂}, whereas this study found relatively low to moderate carbon flux linkages of soil moisture and soil temperature for the smaller (half-hourly) time-scale.

4.4.2 Data quality and uncertainty

Error and uncertainty associated with the eddy fluxes and other variables can cause superfluous biases in analysis and modeling (Williams et al., 2009). The two-step data filtering procedure

substantially eliminated the unsuitable data representing gaps (for more than two variables in one half-hourly panel) and unreasonable spikes (outliers) from the final data sets, which were analyzed to derive and report the results. Furthermore, previous research (Schmidt et al., 2012) reported relatively low measurement uncertainties for canopy-level meteorological variables (relative error $\leq 2\%$), heat fluxes (relative error: 1.7% to 5.2%) and vertical CO₂ fluxes (F_{CO_2}) (relative error $\leq 8.2\%$) of the AmeriFlux network. Schmidt et al. (2012) also suggested high quality research applicability of the AmeriFlux data. Nevertheless, it is possible that the analysis and modeling with half-hourly data may not be completely free from the effect of random sampling errors of measurements (Baldocchi 2003).

4.4.3 Explanatory modeling and analysis

Theoretically process-based carbon dynamics models should be more reliable than the empirical models, but all the relevant processes are not understood yet (Keenan et al., 2011). This reemphasizes the importance of developing data-informed carbon flux modeling system (Keenan et al., 2012), which requires proper mechanistic judgments in the selection of model variables. The multicollinearity effect generally provides biased models with the traditional least squares regression method, which may be partially resolved by eliminating predictors stepwise (backward or forward) or simultaneously based on statistical significance. However, this elimination sometimes results in removal of variables that has important mechanistic basis, hampering the evaluation of comparative linkages of relevant predictors with the response variable. The dimensionless, optimal PLSR modeling approach (capturing maximum system variance and mechanisms) provided the flexibility and statistical stability for retaining all predictors since regression was primarily done with the orthogonal PLS components and then transformed to the original domain. Predictions with the linear PLSR models were quite impressive (R^2 : 0.55 to 0.81), as compared to the nonlinear half-hourly data-driven models of Byrne et al. (2005) for GPP

(R^2 : 0.78 to 0.81) and respiration (R^2 : 0.86 to 0.83), hourly mechanistic model of Wu et al. (2013) for deciduous forest carbon fluxes (R^2 : 0.66 to 0.91), for example.

The PLSR models could not satisfactorily predict extremely positive (upward) and negative (downward) carbon fluxes likely due to the linear structure. Primary production (GPP) and respiration have different mechanisms and are not necessarily controlled by the same set of drivers. Representing these two distinct processes by one of set of regression coefficients at the half-hourly scale can contribute errors to the model predictions. Seasonal variability in terrestrial carbon fluxes can also cause uncertainty in a data-driven model developed with half-hourly data (Jarvis et al., 1997; Xu and Baldocchi 2004). Furthermore, instead of including multiple years of observations for each study site, single-year data, encompassing a 5-year time-frame (2006-11) among the eight deciduous study forests, were chosen for the analysis and modeling. Although a 10-fold cross-validation improved consistency in model-fitting and robustness of estimated parameters, incorporation of single (rather than multiple) year data for an individual site could be seen as a limitation of this study.

Carbon fluxes can strongly respond to precipitation and vegetation productivity (Piao et al., 2013; Schmidt et al., 2011) considering their control over the long-term carbon balance. I used soil moisture (SWC) data as a surrogate for precipitation subject to the lack of availability of half-hourly precipitations. Since the study goal was to determine the relative linkages of mainly climate and environmental variables with the canopy-level vertical carbon fluxes, biological variables such as the canopy leaf-area-index (LAI) was excluded from the predictor data matrix. Further, variables such as SWC and LAI are less likely to change much over a half-hourly interval; the results showed little carbon flux linkages of SWC for the half-hour scale at the US deciduous forests. Although ambient atmospheric concentrations of CO_2 were included in the data matrices, I did not explicitly incorporate anthropogenic carbon sources, which could

influence the large-scale ecosystem carbon emissions. Exclusion of these process components, apart from the linear structure of PLSR, could have contributed to the reduction of the model fitting accuracy (MSE; RSR) and efficiency (R^2) for different study sites.

Building linkages between statistical and mechanistic modeling approaches has been a major challenge in ecological modeling research (Larocque et al., 2011). Issues such as the data requirements, complex parameterizations, prediction uncertainties, computational expenses and expert knowledge basis of available mechanistic carbon-cycle models highlight the importance of developing relatively simple models without conceding the representation of important processes at the relevant spatiotemporal scales. The study presented a simple, systematic multivariate approach to identify the dominant process components by classifying the relevant climate and environmental variables, quantifying their relative linkages with the canopy-level vertical carbon fluxes. The modeling and analysis provides an objective, empirical foundation to obtain crucial mechanistic insights *a priori*; complementing process-based model building with a warranted complexity.

4.5. Conclusions

A data-analytics method was used to determine the relative linkages of different climate and environmental variables with the canopy-level, half-hourly CO₂ fluxes of US deciduous forests. Three biophysical process components were identified to adequately explain the canopy-level vertical CO₂ fluxes. The ‘radiation-energy’ component had the strongest linkages with the canopy-level CO₂ fluxes. The ‘temperature-hydrology’ component showed low to moderate carbon flux linkages. The ‘aerodynamic’ component was relatively weakly connected with the carbon fluxes. The relative linkage of ambient CO₂ concentrations with the vertical carbon fluxes was moderate to strong among different sites. The latent heat flux was the most influential

predictor of instantaneous CO₂ fluxes at all study sites except for Harvard Forest. On average, the ‘radiation-energy’ component showed around 5 and 8 times stronger carbon flux linkages than that of the ‘temperature-hydrology’ and ‘aerodynamic’ components, respectively. The similarity of observed patterns among different study sites (representing sharp gradients in canopy heights, climatic regimes and soil formations) indicates that the findings are potentially transferrable to other deciduous forests around the world. The observed similarities also highlight the scope of developing robust, parsimonious models for appropriate predictions of ecosystem carbon fluxes and potential sequestrations under a changing climate and environment. Relatively good model accuracy and efficiency reiterate the usefulness of multivariate analytics models for gap-filling in time-series of instantaneous flux data.

Future research should investigate the relative linkage patterns by aggregating data in larger time-scales (e.g., daily, weekly, monthly). Multi-scale linkage patterns in other terrestrial ecosystems should also be investigated and compared. More advanced data-analytics approaches such as the system and network modeling, machine learning, and fuzzy logic may also contribute towards developing a robust understanding and prediction of ecosystem carbon fluxes at different spatiotemporal scales.

References

- Akaike, H., 1974. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on* 19(6): 716-723
- AmeriFlux site and data exploration system website, 2014 <http://ameriflux.ornl.gov/>
- Aubinet, M., Grelle, A., Ibrom, A., et al., 1999. Estimates of the annual net carbon and water exchange of forests: the EUROFLUX methodology. *Advances in Ecological Research* 30: 113-175
- Baker, I., Denning, A.S., Hanan, N., et al., 2003. Simulated and observed fluxes of sensible and latent heat and CO₂ at the WLEF-TV tower using SiB2.5. *Global Change Biology* 9(9): 1262-1277
- Baldocchi, D.D., 2003. Assessing the eddy covariance technique for evaluating carbon dioxide exchange rates of ecosystems: past, present and future. *Global Change Biology* 9(4): 479-492
- Baldocchi, D.D., Falge, E., Gu, L., et al., 2001. FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. *Bulletin of the American Meteorological Society* 82(11): 2415-2434
- Baldocchi, D.D., Meyers, T., 1998. On using eco-physiological, micrometeorological and biogeochemical theory to evaluate carbon dioxide, water vapor and trace gas fluxes over vegetation: a perspective. *Agricultural and Forest Meteorology* 90(1-2): 1-25
- Barr, A.G., Black, T.A., Hogg, E.H., Kljun, N., Morgenstern, K., Nesic, Z., 2004. Inter-annual variability in the leaf area index of a boreal aspen-hazelnut forest in relation to net ecosystem production. *Agricultural and Forest Meteorology* 126(3): 237-255
- Beer, C., Reichstein, M., Tomelleri, E., et al., 2010. Terrestrial gross carbon dioxide uptake: global distribution and covariation with climate. *Science* 329: 834-838
- Braswell, B.H., Sacks, W.J., Linder, E., Schimel, D.S., 2005. Estimating diurnal to annual ecosystem parameters by synthesis of a carbon flux model with eddy covariance net ecosystem exchange observations. *Global Change Biology* 11(2): 335-355
- Byrne, K.A., Kiely, G., Leahy, P., 2005. CO₂ fluxes in adjacent new and permanent temperate grasslands. *Agricultural and Forest Meteorology* 135(1): 82-92
- Cao, M., Woodward, F.I., 1998. Dynamic responses of terrestrial ecosystem carbon cycling to global climate change. *Nature* 393(6682): 249-252
- Carrara, A., Kowalski, A.S., Neiryck, J., Janssens, I.A., Yuste, J.C., Ceulemans, R., 2003 Net ecosystem CO₂ exchange of mixed forest in Belgium over 5 years. *Agricultural and Forest Meteorology* 119(3), 209-227

- Carvalhais, N., Seixas, J., Myneni, R., 2005. Modeling net ecosystem productivity scale issues and regional application to the Iberian Peninsula. Seventh International Carbon Dioxide Conference (ICDC7) September 25-30
- Chen, M., Zhuang, Q., Cook, D.R., et al., 2011. Quantification of terrestrial ecosystem carbon dynamics in the conterminous United States combining a process-based biogeochemical model and MODIS and AmeriFlux data. *Biogeosciences Discussions* 8(2): 2721-2773
- Chong, I.G., Jun, C.H., 2005. Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intell Laboratory Systems* 78(1): 103-112
- de Jong, S., 1993. SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intell Laboratory Systems* 18(3): 251-263
- Desai, A.R., 2010. Climatic and phenological controls on coherent regional interannual variability of carbon dioxide flux in a heterogeneous landscape. *Journal of Geophysical Research: Biogeosciences* (2005–2012), 115(G3)
- Dragon, K., 2006. Application of factor analysis to study contamination of a semi-confined aquifer (Wielkopolska Buried Valley aquifer, Poland). *Journal of Hydrology* 331(1): 272-279
- Drake, B.G., Azcon-Bieto, J., Berry, J., et al., 1999. Does elevated atmospheric CO₂ concentration inhibit mitochondrial respiration in green plants? *Plant, Cell & Environment* 22(6): 649-657
- Falge, E., Baldocchi, D., Olson, R., et al., 2001. Gap filling strategies for long term energy flux data sets. *Agricultural and Forest Meteorology* 107(1): 71-77
- Fuentes, J.D., Wang, D., 1999. On the seasonality of isoprene emissions from a mixed temperate forest. *Ecological Applications* 9(4): 1118-1131
- Geider, R.J., Delucia, E.H., Falkowski, P.G., et al., 2001 Primary productivity of planet earth: biological determinants and physical constraints in terrestrial and aquatic habitats. *Global Change Biology* 7(8): 849-882
- Gilmanov, T.G., Parton, W.J., Ojima, D.S., 1997. Testing the 'CENTURY' ecosystem level model on data sets from eight grassland sites in the former USSR representing a wide climatic/soil gradient. *Ecological Modelling* 96(1): 191-210
- Gove, J.H., Holinger, D.Y., 2006. Application of a dual unscented Kalman filter for simultaneous state and parameter estimation in problems of surface-atmospheric exchange. *Journal of Geophysical Research: Atmosphere* 111: 1-21
- Grant, R.F., Baldocchi, D.D., Ma, S., 2012. Ecological controls on net ecosystem productivity of a seasonally dry annual grassland under current and future climates: Modelling with ecosys. *Agricultural and Forest Meteorology* 152: 189-200
- Hargrove, W.W., Hoffman, F.M., 2005 Potential of multivariate quantitative methods for delineation and visualization of ecoregions. *Environmental Management* 34(1): S39-S60

- Haworth, M., Elliott-Kingston, C., McElwain, J.C., 2011. Stomatal control as a driver of plant evolution. *Journal of Experimental Botany* 62(8): 2419-2423
- Heber, U., Neimanis, S., Lange, O.L., 1986. Stomatal aperture, photosynthesis and water fluxes in mesophyll cells as affected by the abscission of leaves. Simultaneous measurements of gas exchange, light scattering and chlorophyll fluorescence. *Planta* 167(4): 554-562
- Heimann, M., Reichstein, M., 2008. Terrestrial ecosystem carbon dynamics and climate feedbacks. *Nature* 451(7176): 289-292
- Hollinger, D.Y., Aber, J., Dail, B., et al., 2004. Spatial and temporal variability in forest-atmosphere CO₂ exchange. *Global Change Biology* 10(10): 1689-1706
- Hubert, M., Branden, K.V., 2003. Robust methods for partial least squares regression. *Journal of Chemometrics* 17(10): 537-549
- Hui, D., Wan, S., Su, B., Katul, G., Monson, R., Luo, Y., 2004. Gap-filling missing data in eddy covariance measurements using multiple imputation (MI) for annual estimations. *Agricultural and Forest Meteorology* 121(1): 93-111
- Jahan, N., Gan, T.Y., 2013. Developing a gross primary production model for coniferous forests of northeastern USA from MODIS data. *International Journal of Applied Earth Observation and Geoinformation* 25: 11-20
- Jarvis, P.G., Massheder, J.M., Hale, S.E., Moncrieff, J.B., Rayment, M., Scott, S.L., 1997. Seasonal variation of carbon dioxide, water vapor, and energy exchanges of a boreal black spruce forest. *Journal of Geophysical Research: Atmospheres* (1984–2012), 102(D24): 28953-28966
- Jolliffe, I.T., 1993 Principal Component analysis: A beginner's guide-II. Pitfalls, myths and extensions. *Weather* 48 (8): 246-253
- Jung, M., Reichstein, M., Margolis, H.A., et al., 2011. Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. *J of Geophysical Res: Biogeosciences* (2005–2012) 116(G3)
- Keenan, T.F., Davidson, E., Moffat, A.M., Munger, W., Richardson, A.D., 2012. Using model-data fusion to interpret past trends, and quantify uncertainties in future projections, of terrestrial ecosystem carbon cycling. *Global Change Biology* 18(8): 2555-2569
- Keenan, T., Maria Serra, J., Lloret, F., Ninyerola, M., Sabate, S., 2011. Predicting the future of forests in the Mediterranean under climate change, with niche-and process-based models: CO₂ matters! *Global Change Biology* 17(1): 565-579
- Kuhn, M., Johnson, K., 2013. *Applied Predictive Modeling*. Springer, NY
- Larocque, G.R., Mailly, D., Yue, T.X. et al., 2011. Common challenges for ecological modelling: Synthesis of facilitated discussions held at the symposia organized for the 2009

- conference of the International Society for Ecological Modelling in Quebec City, Canada, (October 6–9, 2009). *Ecological Modelling* 222(14): 2456-2468
- Launiainen, S., Rinne, J., Pumpanen, J., et al., 2005. Eddy covariance measurements of CO₂ and sensible and latent fluxes during a full year in a boreal pine forest trunk-space. *Boreal Environ Research* 10 (6): 569-588
- Li, Z., Yu, G., Xiao, X., et al., 2007. Modeling gross primary production of alpine ecosystems in the Tibetan Plateau using MODIS images and climate data. *Remote Sensing of Environment* 107(3): 510-519
- Liu, C.W., Lin, K.H., Kuo, Y.M., 2003. Science Application of factor analysis in the assessment of groundwater quality in a blackfoot disease area in Taiwan. *Science of the Total Environment* 313(1): 77-89
- Lloyd, J., Taylor, J.A., 1994. On the temperature dependence of soil respiration. *Functional Ecology* 8: 315-323
- Loescher, H.W., Oberbauer, S.F., Gholz, H.L., Clark, D.B., 2003. Environmental controls on net ecosystem-level carbon exchange and productivity in a Central American tropical wet forest. *Global Change Biology* 9(3): 396-412
- Lund, M., Lafleur, P.M., Roulet, N.T. et al., 2010. Variability in exchange of CO₂ across 12 northern peatland and tundra sites. *Global Change Biology* 16(9): 2436-2448
- Mahbub, P., Ayoko, G.A., Goonetilleke, A., Egodawatta, P., Kokot, S., 2010 Impacts of traffic and rainfall characteristics on heavy metals build-up and wash-off from urban roads. *Environmental Science & Technology* 44(23): 8904-8910
- MÄKELÄ, A., Pulkkinen, M., Kolari, P., et al., 2008. Developing an empirical model of stand GPP with the LUE approach: analysis of eddy covariance data at five contrasting conifer sites in Europe. *Global Change Biology* 14(1): 92-108.
- Masle, J. 2000. The effects of elevated CO₂ concentrations on cell division rates, growth patterns, and blade anatomy in young wheat plants are modulated by factors related to leaf position, vernalization, and genotype. *Plant Physiol* 122(4): 1399-1416
- Melesse, A.M., Hanley, R.S. 2005. Artificial neural network application for multi-ecosystem carbon flux simulation. *Ecological Modelling* 189(3): 305-314
- Moorcroft, P.R., Hurtt, G.C., Pacala, S.W. 2001. A method for scaling vegetation dynamics: the ecosystem demography model (ED). *Ecological Monographs* 71(4): 557-586
- Morales, P., Sykes, M.T., Prentice, I.C. 2005. Comparing and evaluating process-based ecosystem model predictions of carbon and water fluxes in major European forest biomes. *Global Change Biology* 11(12): 2211-2233
- Moriasi, D.N., Arnold, J.G., Liew Van, M.W., Bingner, R.L., Harmel, R.D., Veith, T.L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *American society of Agricultural and Biological Engineers*, 50 (3), 885-900

- Oechel, W.C., Vourlitis, G.L., Verfaillie, J., et al., 2000. A scaling approach for quantifying the net CO₂ flux of the Kuparuk River Basin, Alaska. *Global Change Biology* 6(S1): 160-173
- Panda, U.C., Sundaray, S.K., Rath, P., Nayak, B.B., Bhatta, D. 2006. Application of factor and cluster analysis for characterization of river and estuarine water systems—A case study: Mahanadi River (India). *Journal of Hydrology* 331(3): 434-445
- Peres-Neto, P.R., Jackson, D.A., Somers, K.M., 2003. Giving meaningful interpretation to ordination axes: assessing loading significance in principal component analysis. *Ecology* 84(9): 2347-2363
- Piao, S., Sitch, S., Ciais, P., et al., 2013. Evaluation of terrestrial carbon cycle models for their response to climate variability and to CO₂ trends. *Global Change Biology* 19: 2117-2132
- Post, W.M., Pastor, J., 1996. Linkages—an individual-based forest ecosystem model. *Climatic Change* 34(2): 253-261
- Richardson, A.D., Braswell, B.H., Hollinger, D.Y. et al., 2006. Comparing simple respiration models for eddy flux and dynamic chamber data. *Agricultural and Forest Meteorology* 141(2): 219-234
- Running, S.W., Coughlan, J.C., 1988. A general model of forest ecosystem processes for regional applications I. Hydrologic balance, canopy gas exchange and primary production processes. *Ecological Modeling* 42(2): 125-154
- Running, S.W., Gower, S.T., 1991. FOREST-BGC, a general model of forest ecosystem processes for regional applications. II. Dynamic carbon allocation and nitrogen budgets. *Tree Physiology* 9(1-2) 147-160
- Schimel, D.S., House, J.I., Hibbard, K.A., et al., 2001. Recent patterns and mechanisms of carbon exchange by terrestrial ecosystems. *Nature* 414(6860): 169-172
- Schmidt, A., Hanson, C., Kathilankal, J., Law, B.E., 2011. Classification and assessment of turbulent fluxes above ecosystems in North-America with self-organizing feature map networks. *Agricultural and Forest Meteorology* 151(4): 508-520
- Schmidt, A., Hanson, C., Chan, W.S., Law, B.E., 2012. Empirical assessment of uncertainties of meteorological parameters and turbulent fluxes in the AmeriFlux network. *J of Geophysical Res: Biogeosciences* (2005–2012) 117(G4).
- Schlesinger, W.H., 1999. Carbon sequestration in soils. *Science* 284 (5423): 2095
- Schubert, P., Lagergren, F., Aurela, M., et al., 2012. Modeling GPP in the Nordic forest landscape with MODIS time series data-Comparison with the MODIS GPP product. *Remote Sensing of Environment* 126:136-147
- Sellers, P.J., Dickinson, R.E., Randall, D.A., et al., 1997. Modeling the exchanges of energy, water, and carbon between continents and the atmosphere. *Science* 275(5299): 502-509

- Shir, C.C., Bornstein, R.D. 1977. Eddy exchange coefficients in numerical models of the planetary boundary layer. *Boundary-Layer Meteorology* 11(2): 171-185
- Sims, D.A., Rahman, A.F., Cordova, V.D. et al., 2008. A new model of gross primary productivity for North American ecosystems based solely on the enhanced vegetation index and land surface temperature from MODIS. *Remote Sensing of Environment* 112(4): 1633-1646
- Stauch, V.J., Jarvis, A.J. 2006. A semi-parametric gap-filling model for eddy covariance CO₂ flux time series data. *Global Change Biology* 12 (9): 1707–1716
- White, A., Cannell, M.G., Friend, A.D., 1999. Climate change impacts on ecosystems and the terrestrial carbon sink: a new assessment. *Global Environmental Change* 9: S21-S30
- Williams, M., Richardson, A.D., Reichstein, M., et al., 2009. Improving land surface models with FLUXNET data. *Biogeosciences Discussions* 6(2): 2785-2835
- Wilson, K., Goldstein, A., Falge, E., et al., 2002. Energy balance closure at FLUXNET sites. *Agricultural and Forest Meteorology* 113(1): 223-243
- Wold, H., 1966. Estimation of Principal Components and Related Models by Iterative least squares. In Krishnaiah (ed) *Multivariate Analysis*: 391-420. Academic Press, New York
- Wold, H., 1982. Soft Modelling: The Basic Design and some Extensions. In K.G. Jöreskog & H. Wold (Eds) *Systems Under Indirect Observation, Part II*. Amsterdam: North Holland Press
- Wold, S., Johansson, M., Cocchi, M., 1993. PLS-Partial least squares projections to latent structures. In H. Kubinyi (ed) *3D QSAR in Drug design 1*: 523-550 Kluwer Academic Publishers, Netherlands
- Wold, S., Sjöström, M., Eriksson, L. 2001. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* 58(2): 109-130
- Wu, C., Chen, J.M., 2013. Deriving a new phenological indicator of interannual net carbon exchange in contrasting boreal deciduous and evergreen forests. *Ecological Indicators* 24: 113-119
- Wylie, B.K., Fosnight, E.A., Gilmanov, T.G., Frank, A.B., Morgan, J.A., Haferkamp, M.R., Meyers, T.P. 2007. Adaptive data-driven models for estimating carbon fluxes in the Northern Great Plains. *Remote Sensing of Environment* 106(4): 399-413
- Xu, L., Baldocchi, D.D., 2004. Seasonal variation in carbon dioxide exchange over a Mediterranean annual grassland in California. *Agricultural and Forest Meteorology* 123(1): 79-96
- Zhang, Y., Grant, R.F., Flanagan, L.B., Wang, S., Versegny, D.L. 2005. Modelling CO₂ and energy exchanges in a northern semiarid grassland using the carbon-and nitrogen-coupled Canadian Land Surface Scheme (C-CLASS). *Ecological Modelling* 181(4): 591-6

Chapter 5: Data-analytics for comparative linking of net ecosystem exchange (NEE) with the climatic and ecohydrological variables for diverse ecosystems

Abstract

The relative linkages of net ecosystem exchange (NEE) with the climatic and ecohydrological variables for six diverse AmeriFlux ecosystems were quantified by employing a systematic data-analytics method. Multivariate techniques of correlation matrix, principal component analysis, factor analysis, and explanatory partial least squares regression models were utilized to identify similarity-based groupings of climatic and ecohydrological variables, quantifying their interrelation patterns and comparative linkages with NEE. The analytics with large hourly datasets for the 2002-2012 period indicated a robust pattern of the biophysical process components and their relative linkages with NEE among the six diverse ecosystems, incorporating a large gradient in geographical locations, vegetation and land uses (wetland, grassland, cropland, deciduous, evergreen and mixed forests), hydro-climatic regimes (tropical, temperate, and continental), forest stand age and canopy heights, and soil formations. Four variance-based biophysical components adequately described the overall system-data variation for all six ecosystems. ‘Radiation-energy’ group demonstrated the strongest linkage with NEE; exhibiting 4, 5 and 8 times stronger linkages than that of the ‘temperature-hydrology’, ‘ambient atmospheric CO₂’ and ‘aerodynamic’ components, respectively. Model performance (training $R^2 = 0.57-0.77$, validation $R^2 = 0.58-0.75$) indicated the effectiveness of multivariate analytics to predict NEE fluxes from diverse ecosystems. The observed similarity in latent patterns and relative linkages showed the scope of formulating parsimonious empirical models to obtain robust predictions of ecosystem carbon fluxes at local to regional scales. The data-analytics can be used as an objective empirical framework to achieve crucial mechanistic insights into ecosystem fluxes *a priori*, guiding process-based modeling of a warranted complexity.

5.1 Introduction

Net ecosystem exchange (NEE) represents the net flux of CO₂ between an ecosystem and the atmosphere, indicating the dynamics of ecosystem carbon (Law et al., 2002; Xiao et al., 2008; Tang et al., 2012). NEE accounts for the above-ground assimilatory (photosynthesis by foliage), as well as the above-ground (autotrophic) and below-ground (heterotrophic) respiratory processes. Quantification of the relative linkages of NEE with the relevant climatic and ecohydrological drivers can, therefore, provide important insights and useful information for an appropriate management of ecosystem carbon. Specifically, a quantitative understanding of the comparative linkages can lead to land management strategies and priorities to maximize ecosystem carbon uptake (sequestration) and minimize emissions thereof. Furthermore, transferable patterns of the relative climatic/ecohydrologic linkages of NEE across diverse ecosystems would guide the development of robust models and tools to achieve reliable predictions of ecosystem carbon under a changing climate and environment.

NEE can substantially vary among different ecosystems due to differences in ecosystem functional groups, developmental stages and stand ages, plant heights, disturbance history, and climatic and environmental conditions (Rustad et al., 2001; Wilson et al., 2001; Thornton et al., 2002; Litvak et al., 2003). In particular, the perplexing associations between NEE and its physical drivers impede an integrated, quantitative assessment of their relative linkages (Stoy et al., 2005). The vegetation disparity can also contribute to site-specific discrepancies in environmental drivers of NEE (Goulden et al., 1996; Law et al., 2003). Given the multi-faceted associations and interactions of NEE processes, a systematic multivariate data-analytics framework can help to identify the dominant process variables and quantify their relative linkages with NEE; exploring similarity and robust patterns among diverse ecosystems.

Previous research (e.g., Archibald et al., 2009; Reverter et al., 2010) suggested solar radiation, air and soil temperatures as the major drivers of NEE. Loescher et al. (2003) reported a very strong correlation ($R^2 = 0.98$) between the nocturnal NEE and the corresponding friction velocity (UST) for a tropical wet forest; however, the day-time UST and NEE were not notably correlated. Loescher et al. also found a moderate correlation ($R^2 = 0.51$) between NEE and photosynthetic active flux density, while weak correlations with air temperature ($R^2 = 0.18$) and vapor pressure deficit (VPD) ($R^2 = 0.21$). Otieno et al. (2012) investigated the sensitivities of NEE for peatland herbaceous vegetation in response to variation of VPD, peat temperature, ground-water level and stomatal conductance. They largely attributed the low rate of NEE to the stomatal closure and low light utilization efficiency (decreased photosynthetic capacity) of the vegetation subject to high VPD during summer.

Various process-based models have been developed to simulate and predict ecosystem CO₂ exchanges by mechanistically describing different process components. Process-based models are generally developed based on preconceived scientific hypotheses, and the outputs are inherently tied to the embedded hypotheses, model structure, and parameterizations (Beer et al., 2010). They are typically more suitable for the longer time-scales, requiring a broad range of data, methodologies and ecological process descriptions (Keenan, 2012). Larocque et al. (2008) expressed concerns over the inadequate availability of mechanistic model functions for fully describing the carbon cycle processes, emphasizing the accurate evaluation of natural variability and uncertainty estimation of the associated parameters.

Examples of process-based carbon dynamics models are ecosys (Grant et al., 2012), IFUSE (Desai, 2012), DLEM (Tian et al., 2010), ED2 (Medvigy et al., 2009), SiB3 (Baker et al., 2008), CT (Peters et al., 2007), ORCHIDEE (Krinner et al., 2005), Forest-DNDC (Li and Cui, 2004), Biome-BGC (Thornton et al., 2002), ISOLSM (Riley et al., 2002), IBIS (Foley et al.,

1996; Kucharik et al., 2000), HYBRID (White et al., 1999), CENTURY (Glimanov et al., 1997), LINKAGE (Post and Pastor, 1996), and EBL (Shir and Bornstein, 1977). Despite involving detailed process descriptions at varying levels, the mechanistic models cannot often predict ecosystem carbon exchanges with much certainty. Schwalm et al. (2010) compared 22 terrestrial biosphere models to predict monthly CO₂ exchanges for 44 eddy covariance sites, and reported an overall poor performance of the models and high uncertainty in the existing parameter sets.

Following the advent of the current data acquisition and computation era, ecological research is quickly moving toward a ‘rich data’ paradigm. Significant advancements have been made in building public databases (e.g., AmeriFlux; see Baldocchi, 2003) for ecosystem carbon-water-energy fluxes and the associated biophysical variables. Data-analytics can be a crucial vehicle to unveil the latent and useful information from the plethora of ‘raw ecological data’. The advantage of data-analytics based empirical approaches is that they do not necessarily rely on preconceived hypotheses or assumptions and, therefore, do not require any governing mechanistic equations. Empirical models can be relatively simple in structure, parsimonious (involving minimum parameters), and computationally inexpensive; often building the necessary foundation to develop appropriate mechanistic models with a warranted complexity.

Linear and non-linear regressions, artificial neural network, and probabilistic models have already been successfully used for gap-filling and estimations of ecosystem carbon, water and energy fluxes (e.g., Lloyd and Taylor, 1994; Oechel et al. 2000; Falge et al., 2001; Papale and Valentini, 2003; Carrara et al. 2003; Hollinger et al., 2004; Hui et al., 2004; Braswell et al., 2005; Byrne et al. 2005; Gove and Hollinger, 2006; Richardson et al., 2006; Stauch and Jarvis, 2006; Wylie et al., 2007; Mäkelä et al., 2008; Xiao et al., 2008; Jahan and Gan, 2013; Menzer et al., 2013). However, existing data-driven models are mostly limited to the at-site scale; the site-specific parameterizations are often not transferable to another similar, let alone diverse,

ecosystem site. Further, a quantitative understanding on the possible general set of dominant predictors, their complex interactions, and relative linkages with NEE across diverse ecosystems is yet to be achieved. Filling this information-gap on ecological similitude is crucial for developing at-site (ecosystem) to regional-scale parsimonious empirical models to achieve robust predictions of ecosystem carbon fluxes.

Ishtiaq and Abdul-Aziz (2015) developed a multivariate data-analytics and informatics approach to determine the relative linkages of canopy-level, half-hourly CO₂ fluxes with the climatic and environmental variables for eight US deciduous forests. The study reported robust carbon flux linkage patterns across different deciduous forests, representing geographical gradients in climate, canopy heights, topography, hydrology, and soil-formations. On average, the vertical CO₂ fluxes showed substantially (5-8 times) stronger linkages with the radiation-energy component (representing radiation and heat fluxes) than that with the temperature-hydrology (e.g., VPD, air and soil temperatures, soil moisture) and aerodynamic (e.g., wind speed, friction velocity) components.

The scientific objective of this paper is to quantify interrelations and comparative linkages of NEE with the relevant climatic and ecohydrological variables for diverse ecosystems, incorporating a large gradient in geographical locations, vegetation and land uses (wetland, grassland, cropland, deciduous, evergreen and mixed forests), hydro-climatic regimes (tropical, temperate, and continental), forest stand age and canopy heights, and soil formations. The paper builds on the study of Ishtiaq and Abdul-Aziz (2015) by employing the systematic data-analytics framework with large hourly data matrices for six diverse AmeriFlux ecosystems. Multivariate pattern recognition techniques of correlation matrix, principal component and factor analyses were utilized to discern the interrelations, relative influence, similarity, and grouping patterns of the climatic, environmental and ecological variables for different ecosystems. Dimensionless (i.e.,

standardized) partial least squares regression (PLSR) models were then developed to quantify the relative linkages (weights) of different climatic and environmental variables with NEE. The findings will enrich the current understanding of carbon flux dynamics for diverse ecosystems, and guide the development of parsimonious empirical to detailed process-based models of a warranted complexity to achieve robust predictions of NEE at variable scales.

5.2 Materials and methods

5.2.1 Diverse ecosystem case study sites

Six different ecosystems were chosen as case studies from the AmeriFlux network (<http://ameriflux.ornl.gov>) to represent distinct climate, vegetation and land cover types based on the International Geosphere-Biosphere Programme for land cover classification (Belward, 1996). The selected study sites are as follows (Figure 5.1, Table 5.1): (i) Sylvania Wilderness, Michigan (mixed forest); (ii) Fermi Prairie, Illinois (grassland); (iii) Mead Irrigated, Nebraska (croplands); (iv) Niwot Ridge, Colorado (evergreen needleleaf forest); (v) Morgan Monroe State Forest, Indiana (deciduous broadleaf forest); and (vi) Florida Everglades Taylor Slough Short Hydroperiod Marsh, Florida (permanent, coastal wetlands). The study ecosystems, therefore, represented an extensive variation in plant characteristics (e.g., canopy heights, leaf area, root zones), photosynthesis-respiration, and associated carbon fluxes. Further, the large geographical region represented different climate (e.g., continental vs. temperate vs. tropical), land uses (vegetation), hydrology, soil geomorphology, and terrain patterns; incorporating a spatial gradient in land-atmospheric NEE fluxes.

The Sylvania Wilderness (canopy height: 22 m) is one of the oldest US forests with an average stand age of 350 years. It has a warm summer continental climate with a short, moist growing season and relatively cold winter; its soil is moderately drained coarse or sandy loam

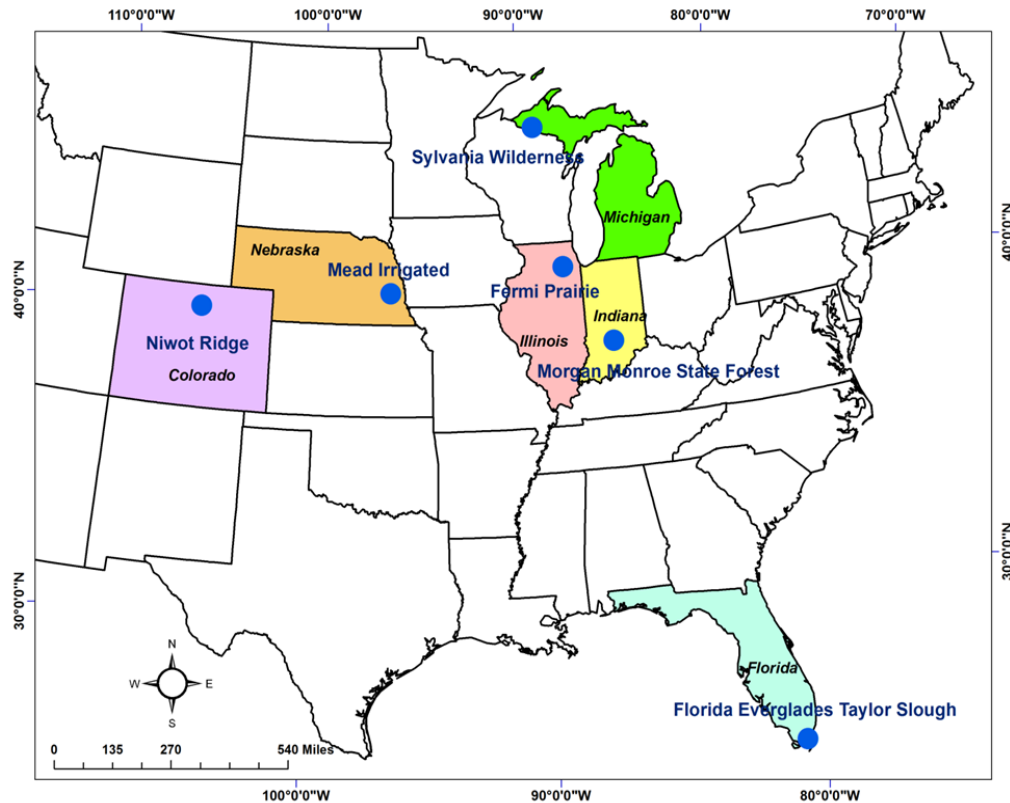


Figure 5.1: Locations of six diverse AmeriFlux ecosystem study sites.

spodosols (ashy gray, acidic soils). The Mead Irrigated and Fermi Prairie share a humid continental climate. The Mead Irrigated (plant height: 2.9 m) has mostly been used as an agricultural land; it is characterized by a deep silty, clay loam soil (typical for eastern Nebraska). The Fermi Prairie has tall prairie grasses since 1989; the topsoil is silty clay loam, while the subsoil contains clay. The Niwot Ridge (canopy height: 11.5 m) has a subarctic climate and an average forest stand age of 102 years; its soil is characterized by extremely rocky (granite) mineral clays. The Morgan Forest (canopy height: 27 m) is an old forest having an average stand age of 70 years, with a humid subtropical climate and mesic typic, Dystrochrept, clay soil. The Florida Everglades Taylor Slough is a sub-tropical, coastal, permanent freshwater wetland with a short hydroperiod; the climate is represented by tropical wet and dry seasons. The oligotrophic marsh is characterized by marl (calcium carbonate) soils, whereas the plant community is

Table 5.1: Bio-physical summary of the diverse ecosystem study sites from the AmeriFlux network

Ecosystems	Vegetation types	Latitude (°N)	Longitude (°W)	Canopy height (m)	Koppen climate classification	Mean annual temperature (°C)	Mean annual precipitation (mm)	References
Sylvania Wilderness (MI)	Mixed forest	46.24	89.35	22	Dfb (Warm summer continental)	3.81	826	Desai et al., 2005
Fermi Prairie (IL)	Grassland	41.84	88.24		Dfa (Humid continental)	9.04	930	Matamala et al., 2008
Mead Irrigated (NE)	Cropland	41.17	96.48	2.9	Dfa (Humid continental)	10.07	790	Verma et al., 2005; Suyker et al., 2004, 2005
Niwot Ridge (CO)	Evergreen needleleaf forest	40.03	105.55	11.5	Dfc (Subartic)	1.50	800	Monson et al., 2005, 2002
Morgan Monroe State Forest (IN)	Deciduous broadleaf forest	39.32	86.41	27	Cfa (Humid subtropical)	10.85	1032	Dragoni et al., 2010; Oliphant et al., 2004
Florida Everglades Taylor Slough Short Hydroperiod Marsh (FL)	Wetland	25.44	80.59		Am (Tropical)	23.9	1430	Schedlbauer et al., 2010; 2012

dominated by a relatively uniform grass-sedge canopy (Davis et al., 2005; Schedlbauer et al., 2012).

5.2.2 The AmeriFlux data sets

Eddy covariance-based hourly measurements of NEE and the corresponding climatic and ecohydrologic variables were obtained from the AmeriFlux network (<http://ameriflux.ornl.gov>) for different five years (2002-2006; 2005-2009; 2007-2011; 2008-12) during 2002-2012 for the six different study ecosystems. The 5-year periods for each ecosystem and the overall 11-year period among different ecosystems incorporated the multi-scale (e.g., hourly, daily, seasonal, inter-annual, and decadal) temporal variabilities of NEE and relevant climate and ecohydrologic processes (see Table 5.2 for data summary and statistics). Selected variables were NEE (including sub-canopy CO₂ storage); ambient atmospheric CO₂ concentration (C_{CO2}); net radiation (RN); photosynthetically active radiation (PAR); sensible heat flux (SHF); latent heat flux (LHF) (ecosystem water exchange); soil heat flux (FG); air temperature (TA); soil temperature (TS); vapor pressure deficit (VPD); soil water content (SWC) (unavailable for the Mead Irrigated site); wind speed (WS); and friction velocity (UST). Overall, the observational data matrix included 12 potential explanatory variables of NEE for an individual ecosystem. The participating variables were chosen based on a preliminary data analysis and the existing knowledge of ecosystem carbon flux dynamics. According to the AmeriFlux sign conventions, positive sign represented upward fluxes (land/forest to atmosphere) of NEE, LHF, and SHF; negative refers to their downward fluxes (atmosphere to land/forest). In contrast, positive values of radiation (RN and PAR) and FG indicate downward fluxes and vice versa.

This study utilized the level-2 AmeriFlux data that had passed through the preliminary QA/QC checks. Given the very large sample sizes (N = 26660-41999) for different ecosystems representing different periods of five consecutive years, gaps in the primary data sets

Table 5.2: Summary of data records and statistics (averages and standard deviations) for the observed hourly NEE, climate and ecohydrological variables of the six study sites. Standard deviations are given in parenthesis.

Ecosystem sites	Data Records	RN (W/m ²)	PAR (μmol/m ² /s)	LHF (W/m ²)	SHF (W/m ²)	FG (W/m ²)	TA (°C)	TS (°C)	VPD (KPa)	SWC (%)	WS (m/s)	UST (m/s)	C _{CO2} (μmol/mol)	NEE (μmol/m ² /s)
Sylvania Wilderness	2002-06 <i>N</i> =24475	64.98 (172.97)	323.38 (516.25)	29.60 (62.03)	18.53 (76.02)	0.32 (17.63)	5.76 (12.08)	7.97 (7.36)	0.41 (0.49)	20.34 (3.22)	3.15 (1.16)	0.54 (0.28)	380.75 (11.38)	-0.31 (5.02)
Fermi Prairie	2007-11 <i>N</i> =29131	106.93 (189.76)	389.87 (534.13)	58.15 (91.36)	20.56 (66.69)	2.39 (36.89)	9.12 (12.04)	12.32 (7.96)	0.51 (0.53)	38.47 (6.67)	2.88 (1.63)	0.31 (0.18)	376.72 (26.75)	-1.44 (6.38)
Mead Irrigated	2008-12 <i>N</i> =30787	111.45 (218.92)	407.20 (564.02)	60.59 (103.96)	22.55 (71.56)	3.86 (50.12)	10.81 (12.00)	11.57 (10.1)	0.53 (0.62)		3.96 (2.41)	0.29 (0.19)	399.97 (29.57)	-1.62 (11.83)
Niwot Ridge	2008-12 <i>N</i> =35592	100.07 (236.60)	351.11 (525.42)	48.09 (63.62)	40.04 (140.3)	-0.41 (5.67)	2.36 (9.08)	2.56 (4.10)	0.46 (0.41)	9.88 (7.09)	4.80 (3.40)	0.75 (0.58)	391.37 (4.52)	-0.69 (3.74)
Morgan Forest	2005-09 <i>N</i> =23422	146.94 (226.58)	485.89 (585.00)	55.91 (95.53)	37.70 (90.59)	0.49 (9.60)	11.84 (10.87)	11.06 (6.76)	0.61 (0.57)	37.43 (7.25)	3.80 (1.51)	0.57 (0.26)	392.04 (12.81)	-2.73 (7.90)
Florida Everglades	2008-12 <i>N</i> =16457	251.75 (269.93)	709.46 (637.66)	96.42 (73.17)	71.57 (97.41)	9.45 (32.18)	25.31 (5.80)	23.82 (4.50)	1.05 (0.60)	68.05 (15.9)	3.19 (1.30)	0.35 (0.13)	388.45 (12.08)	-0.93 1.78

Notes: ‘Blank’ in the table refer to ‘missing data’. *N* = sample size. RN, PAR, LHF, SHF, FG, TA, TS, VPD, SWC, WS, UST, C_{CO2}, and NEE, respectively, refer to net radiation, photosynthetically active radiation, latent heat flux, sensible heat flux, soil heat flux, air temperature, soil temperature, vapor pressure deficit, soil water content, wind speed, friction velocity, ambient atmospheric CO₂ concentration, and net ecosystem exchange.

Table 5.3: Pearson correlation coefficients between the NEE and climatic/ecohydrological variables for the diverse ecosystem study sites

Ecosystem sites	RN	PAR	LHF	SHF	FG	TA	TS	VPD	SWC	WS	UST	C _{CO2}
Fermi Prairie	-0.74	-0.73	-0.84	-0.33	-0.54	-0.34	-0.25	-0.49	0.16	-0.09	-0.15	0.39
Sylvania Wilderness	-0.69	-0.68	-0.70	-0.54	-0.48	-0.22	-0.18	-0.39	0.05	-0.01	-0.15	0.39
Niwot Ridge	-0.67	-0.67	-0.66	-0.53	-0.52	-0.33	-0.24	-0.34	-0.19	0.10	0.01	0.34
Florida Everglades	-0.69	-0.71	-0.61	-0.76	-0.25	-0.36	0.01	-0.51	0.14	-0.28	-0.45	0.22
Mead Irrigated	-0.60	-0.56	-0.78	-0.15	-0.30	-0.31	-0.26	-0.35		0.04	-0.21	0.41
Morgan Forest	-0.72	-0.71	-0.85	-0.32	-0.39	-0.52	-0.52	-0.49	0.23	0.36	0.19	0.41

Notes: Blank indicates missing data

were not filled to avoid artificial biases in the data-analytics outcomes. To remove outliers and unsuitable data from each site, we applied a two-layer filtering approach following Ishtiaq and Abdul-Aziz (2015). The first filter removed hourly data panels having gaps for more than two participatory variables; the gap-filtering led to exclusion of hourly panel data by around 18% for Niwot Ridge, by 30% for Mead Irrigated, by 33% for Fermi Prairie, by 44% for Sylvania Wilderness, by 46% for Morgan forest, and by 62% for Florida Everglades from their respective primary data sets of 5-year periods. The removal percentages were higher for the latter three stations, compared to the 35% missing data encountered annually on average among different ecosystems at the half-hourly scale (see Falge et al., 2001). Apart from inclement weather conditions and instrument (eddy covariance tower equipment) malfunctioning, gaps in the hourly data panels were contributed by missing data for ecohydrologic variables such as soil moisture (SWC) that had not been measured by the eddy covariance technique. In the second filtering step, gap-filtered data of each variable for each ecosystem were plotted with time (not shown) to visually identify the unreasonable spikes (i.e., outliers), which were excluded from further analysis. This second filtering removed no hourly observation panels from the Florida Everglades and Morgan Forest, while removing 4 hourly observation panels from the Niwot Ridge, 6 observation panels from the Fermi Prairie, 12 observation panels from the Mead Irrigated, and 20 observation panels from the Sylvania Wilderness. Encompassing a large geographical extent over the 11 year period (2002-12), the final data sets for diverse ecosystems (N= 16457-35592; see Table 5.2) incorporated the effects of relatively large spatiotemporal gradients and variability of climatic, hydrological, biological, and ecological processes on NEE fluxes.

Overall, the data sets included atmospheric, hydrological, and ecological variables controlling NEE (photosynthesis and respiration) at different levels; e.g., from the leaf-level gross primary production (GPP) to the ecosystem-level autotrophic and heterotrophic respiration. On average among different ecosystems, the final data sets represented positive and negative NEE

equally (i.e., by around 50% for each flux type). Biological variables (e.g., biomass, leaf-area-index) were not included in the data-analytics since the objective of this study was to determine the relative linkages of mainly climatic and environmental variables with the NEE. Subject to the lack of precipitation data at the hourly scale, SWC was considered to represent the hydrologic forcing on the NEE processes. Further, the partitioned heat fluxes (LHF, SHF, and FG) were included in the analysis to incorporate the ecohydrologic forcing of NEE.

5.2.3 The data-analytics framework

The data-analytics (Figure 5.2), as employed in this study, involved a systematic application of Pearson product-moment correlation analysis, unsupervised principal component analysis (PCA) (Jolliffe I. T., 1993; Peres-Neto et al., 2003; Mahbub et al., 2010), maximum likelihood estimation based factor analysis (FA) (Panda et al, 2006; Dragon, 2006; and Liu et al., 2003), and orthogonal component based partial least squares regression (PLSR) modeling (Wold 1966, 1982). For PCA, FA and PLSR modeling, data for all the variables were first transformed to the respective Z-scores (i.e., $Z = (X - \mu)/\sigma$, where $Z = Z\text{-score} = \text{standardized variable}$, $X = \text{original variable}$, $\mu = \text{site-specific mean of } X \text{ over the study period}$, and $\sigma = \text{standard deviation of } X$) in order to bring different units of variables to a comparable reference scale. Pearson correlation matrix indicated the linear correspondence between NEE and the climate/ecohydrological variables, as well as the multicollinear structure of the data. PCA and FA identified the hidden grouping and inter-relation patterns of explanatory variables, indicating their relative linkages with NEE. The PLSR models with Z-scores, upon transformation from the orthogonal component domain to the original variables, quantified the relative linkages of NEE with the different climatic and ecohydrological drivers for the respective study ecosystems.

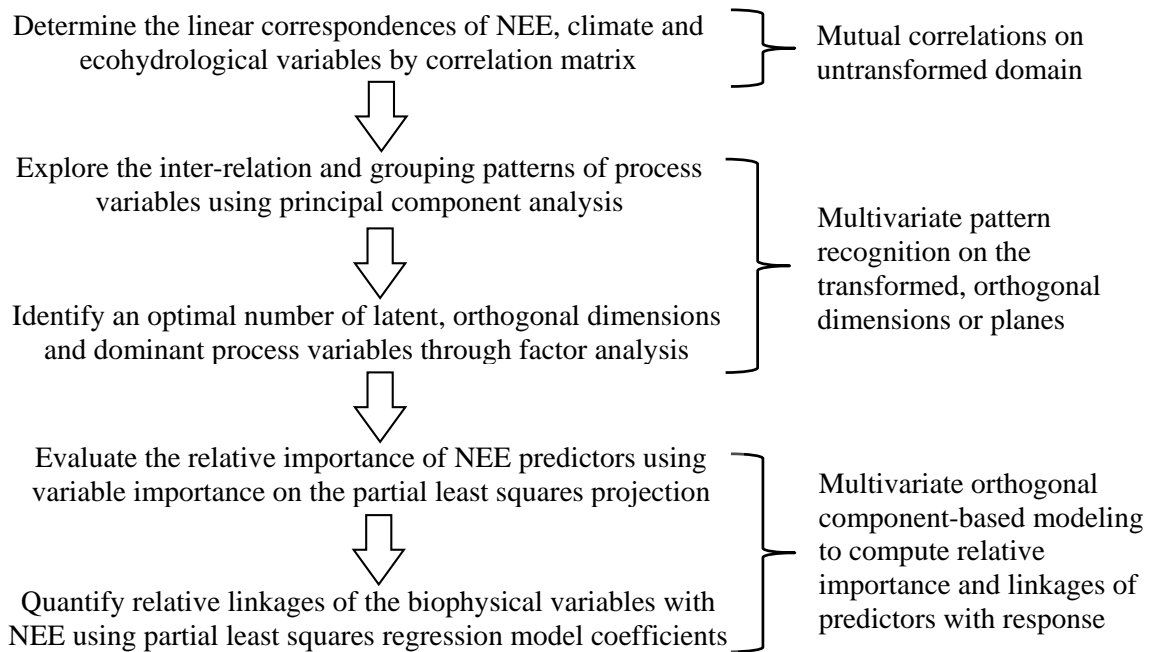


Figure 5.2: The data-analytics framework

Data for the respective five years (i.e., 2002-2006, 2005-2009, 2007-2011, and 2008-12) for different ecosystems were used for correlation analysis, PCA, and FA. The advantage of PCA is that it can extract hidden multivariate patterns from the original data in an unsupervised manner by deriving some orthogonal entities called principal components (PCs), which are linear combinations of the original variables. The orthogonalization ensures representation of the multivariate data sets on uncorrelated planes. Typically, the first two PCs are extracted from a loading matrix (indicating correlations between the PCs and the original variables) and represented through biplots to visualize any hidden grouping and interrelations (orientations and correlation lengths) among the participatory variables. In contrast, FA can be considered a supervised PCA, converting the original variables into a set of rudimental but unobservable orthogonal quantities called factors. The maximum likelihood estimation of a factor loading matrix maximizes the total system variance explained, while the ‘varimax’ factor-rotation

optimizes the loadings (i.e., correlation coefficients) of participatory variables on each factor. Individual latent factors are then extracted based on the eigenvalue criterion (eigenvalue > 1.0) to represent the most system variance of a data matrix with a minimum number of factors.

The PLSR modeling is performed on the transformed orthogonal planes by fitting the independent partial least squares (PLS) components (rather than the original predictor variables) with the response variable, and by maximally linking the predictor-data covariance with that of the response. The issue of multicollinearity among the original predictors (and unreliable predictions therefrom) is thus resolved in the orthogonal domain, allowing the retention of all original predictors in the final model. The regression coefficients of the original, independent variables are then obtained from that of the optimal PLSR model by leveraging the linear transformations between the PLS components and the original variables (Kuhn and Johnson, 2013).

The site-specific PLSR models were trained (i.e., calibrated) for each ecosystem with the corresponding data of the first four years (i.e., Fermi Prairie: 2007-2010; Morgan State Forest: 2005-2008; Sylvania Wilderness: 2002-2005; Mead Irrigated, Niwot Ridge and Florida Everglades: 2008-2011). Data for the remaining single years (Fermi Prairie: 2011; Morgan State Forest: 2009; Sylvania Wilderness: 2006; Mead Irrigated, Niwot Ridge and Florida Everglades: 2012) were used for model testing and validations. The SIMPLS algorithm (de Jong, 1993) along with a 10-fold cross-validation (Kuhn and Johnson, 2013) was applied to achieve a consistent model fitting and robust estimation of parameters for the 4-year calibration periods. A modified Akaike Information Criterion (AIC) (Akaike, 1974; Abdul-Aziz and Ishtiaq, 2014) and the coefficient of determination (R^2) were used to determine the minimum number of PLS components required to obtain the optimal PLSR model based on the minimum prediction error (minimum AIC) and maximum explained variance of NEE (maximum R^2). The modified,

normalized AIC_{NEE} was computed to bring data for different ecosystems (having different sample sizes) to a comparable scale as: $AIC_{NEE}(p) = \ln(SSE_{NEE}/N) + 2p/N$; where p is the total number of the PLS components, N is the sample size, and SSE_{NEE} is the total sum of squared errors between the modeled NEE (Z-scores) and the corresponding observations. Apart from the model efficiency metric (R^2), ratio of root-mean-square error to the standard deviation of observations (RSR) and mean square error (MSE) were used to assess accuracy of the final models. RSR was computed as: $RSR = \sqrt{\sum_{i=1}^N (NEE_{i,mod} - NEE_{i,obs})^2 / N} / \sigma_{NEE,obs}$, where $NEE_{i,obs}$ and $NEE_{i,mod}$ are the i^{th} observed and predicted (respectively) Z-scores of NEE, $\sigma_{NEE,obs} = 1.0$ is the standard deviation of the observed Z-scores of NEE, and N is the sample size. Following Moriasi et al. (2007), a perfect to very good model has an RSR between 0 and 0.50, a good model has an RSR between 0.50 and 0.60, and a satisfactory model has an RSR between 0.60 and 0.70; $RSR > 0.70$ indicates an unsatisfactory model.

Since the PLSR model was developed with Z-score variables (leading to a ‘zero’ model intercept), the model coefficients (BETA), along with the ‘variable importance in the projection (VIP)’ scores (Wold et al., 1993, 2001; Chong and Jun, 2005), were used to determine the relative contributions and linkages of different predictors (PAR, RN, LHF, SHF, FG, VPD, TA, TS, SWC, C_{CO2}, UST and WS) to the response variable (NEE). Higher VIP scores and BETA coefficients indicated the stronger predictors of NEE in the latent predictor-response matrix; $VIP > 1.0$ referred to the most informative predictors (Kuhn and Johnson 2013). The dimensionless PLSR model coefficients were also aggregated to quantify the overall relative linkages (see section 3.4) of different process components (e.g., radiation, heat fluxes, temperature, hydrology, aerodynamic) with the NEE for the different study ecosystems.

5.3 Results

5.3.1 Linear correspondences of NEE, climatic and ecohydrological variables

The statistically significant, Pearson correlation coefficients ($\alpha = 0.05$ for a two-tailed test) between hourly NEE and the corresponding climatic and ecohydrological variables were reported for the six study ecosystems (Table 5.3). Although correlation analysis was performed with untransformed (i.e., non-standardized) data of the original variables, only absolute values of the correlation coefficients were used to illustrate linear correspondences among the participatory variables given that NEE, heat fluxes, net radiation, air and soil temperatures included both positive and negative observations. For all ecosystems, the radiation variables (RN and PAR) demonstrated relatively high linear correspondences ($|r| = 0.56-0.74$) with NEE. Among the heat fluxes (SHF, LHF, and FG), LHF (ecosystem-atmospheric heat and water exchanges due to evapotranspiration) were relatively highly correlated ($|r| = 0.61-0.85$) with NEE. SHF (ecosystem-atmospheric heat exchanges due to temperature gradient) and FG (soil heat flux) showed relatively moderate correlations ($|r| = 0.15-0.54$) with NEE for all but one ecosystem; NEE was strongly correlated ($|r| = 0.76$) with SHF for the Florida Everglades. Relatively moderate linear correspondences of NEE were observed with TA (air temperature, $|r| = 0.22-0.52$), TS (soil temperature, $|r| = 0.18-0.52$), VPD (vapor pressure deficit, $|r| = 0.34-0.51$), and C_{CO_2} (ambient CO_2 concentration, $|r| = 0.22-0.41$) at all ecosystems except for the Florida Everglades, where TS was very weakly correlated with NEE ($|r| = 0.01$). Relatively weak linear correspondences of hourly NEE were generally noted with the aerodynamic drivers (UST and WS; $|r| = 0.01-0.21$) and with the hydrological variable (SWC; $|r| = 0.05-0.23$) with some exceptions for Morgan Forest and the Florida Everglades. Wind speed (WS) showed a moderate correlation ($|r| = 0.36$) with NEE at Morgan Forest, while both WS and friction velocity (UST) were moderately correlated ($|r| = 0.28$ and 0.45 , respectively) with NEE at the Everglades site;

indicating a relatively stronger influence of boundary layer mixing on carbon fluxes in the coastal wetland.

The triangular correlation matrix (not shown) for each study site indicated high mutual correlations ($|r| = 0.82-0.98$) among the radiation and heat flux variables (SHF, LHF, FG, RN and PAR). Mutual correlations ($|r| = 0.73-0.94$) among the temperature variables (TS, TA and VPD), as well as that between the aerodynamic variables (WS and UST; $|r| = 0.72-0.95$) were also notably high. The high mutual correlations indicated the presence of a considerable multicollinearity in the climate and ecohydrological data matrices.

5.3.2 Relative orientations and groupings of NEE and driving variables

The PCA loadings (correlation coefficients between the PCs and the original variables) for the different study ecosystems are presented through biplots (Figure 5.3). Among the six diverse ecosystems, the first two PCs explained 87% to 91% of the total data-system variance.

The interrelation and hidden grouping patterns of the participatory variables were similar for Fermi Prairie (grassland; Figure 5.3a), Sylvania Wilderness (mixed forest; Figure 5.3b), and Niwot Ridge (evergreen needle leaf forest; Figure 5.3c) ecosystem sites. In Fermi Prairie, for example, the relative orientations, loadings and correlation lengths of RN, PAR, LHF, SHF and FG suggested high interrelationships, forming a dominant group (A) that had strong linkages (highly non-orthogonal orientations) with NEE and ambient C_{CO_2} (Figure 5.3a). The TA, TS and VPD formed a second group (B) that showed a moderately non-orthogonal orientation and moderate linkage with NEE despite the relatively long correlation lengths and loadings on both principal components (PCs). The opposite orientation (180°) between SWC and Group-B variables indicated their notably negative mutual correlations. The UST and WS formed a third group (C), exhibiting a nearly moderate non-orthogonality with NEE and long correlation lengths

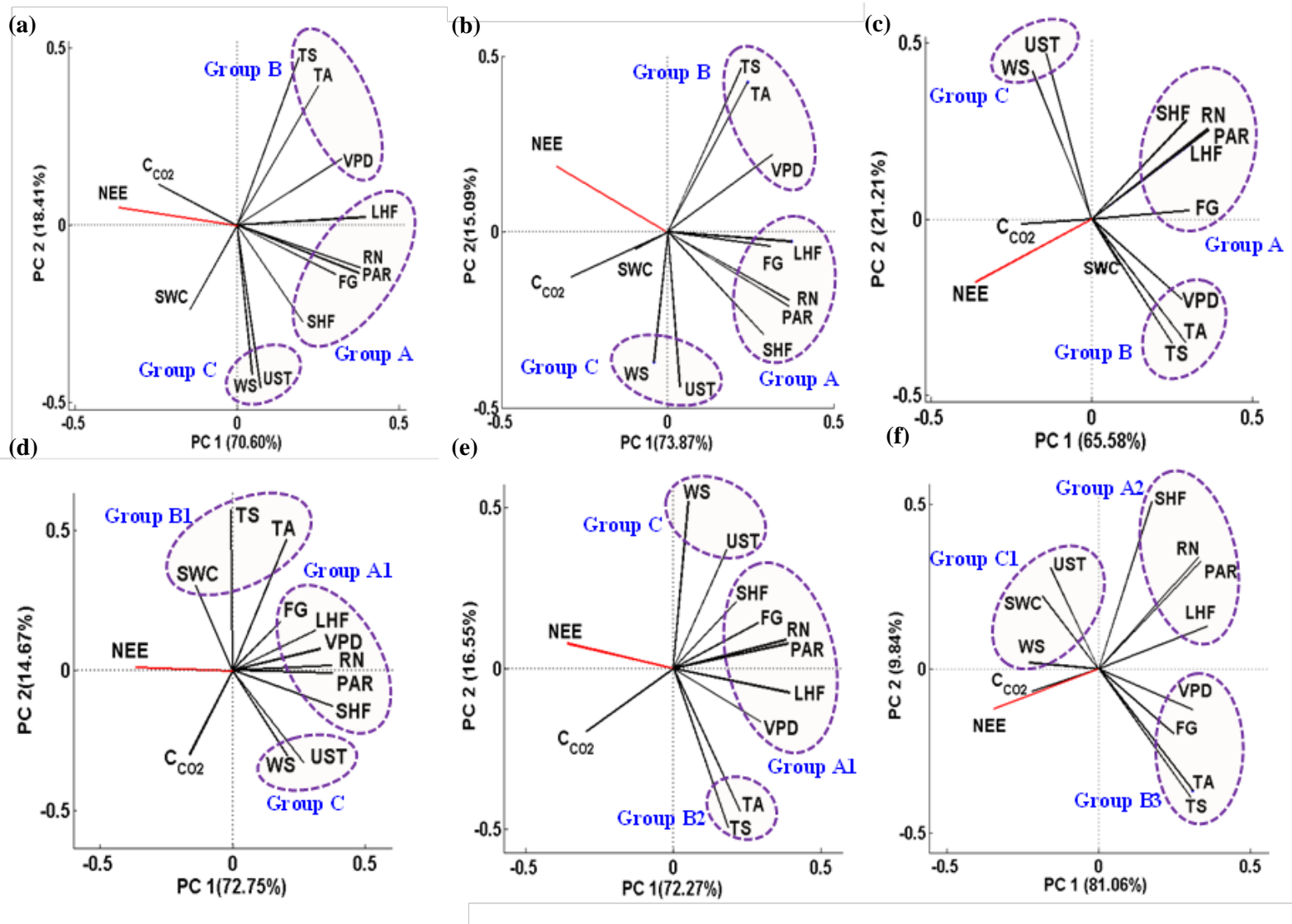


Figure 5.3: Biplots from principal component analysis, showing the interrelation and grouping patterns of climatic, ecohydrological and biological variables for (a) Fermi Prairie (IL), (b) Sylvania Wilderness (MI), (c) Niwot Ridge Forest (CO), (d) Everglades (FL), (e) Mead Irrigated (NE), and (f) Morgan Monroe State Forest (IN). Percent variance explained by each PC is shown in parentheses.

similar to that of Group-B; however, the substantially small loadings of both WS and UST on PC1 suggested a relatively weak linkages of Group-C variables with NEE.

Although the interrelation, grouping and NEE linkage patterns for the other three ecosystems (Florida Everglades, Mead Irrigated, and Morgan State Forest) were nearly similar to that of Fermi Prairie, some important exceptions were evident. For the Florida Everglades (coastal wetland) and Mead Irrigated (cropland) sites, VPD emerged out of the group of temperature variables (Group-B1 and B2) and loaded on Group-A1 with the radiation-heat flux variables (see Figure 5.3d,e). Further, the SWC (missing for the Mead Irrigated site) loaded with the temperature variables (group-B1) at the Everglades site. However, unlike other study ecosystems, both Group-B1 (or B2) and Group-C appeared nearly equally (moderately) linked with the hourly NEE for the Everglades and Mead Irrigated sites. For Morgan State Forest (deciduous broadleaf forest), soil heat flux (FG) emerged out of the radiation-heat fluxes cluster (Group-A2) and loaded with the temperature variables (TA, TS and VPD; Group-B3) (Figure 5.3f); although SWC grouped with the aerodynamic variables (UST and WS; Group-C1), strongly negative correlations of SWC with the temperature variables (Group-B3) were evident. However, similar to Fermi Prairie, the Group-B3 and Group-C1 appeared, respectively, moderately and weakly linked with NEE for the Morgan Forest site.

In general, PCA identified three variance-based groups hidden in the data matrices; Group-A (or A1-A2) was dominated by the radiation and heat flux variables; Group-B (or B1-B3) was dominated by the temperature variables; and Group-C (or C1) was dominated by the

aerodynamic variables. For all six ecosystems, the near-canopy atmospheric carbon dioxide concentration (C_{CO_2}) emanated alongside NEE with small angles. However, relatively small correlation lengths of C_{CO_2} on the biplots, compared to the other variables, indicated its moderate linkage with hourly NEE.

5.3.3 Dominant biophysical variables based on optimal, orthogonal latent factors

The ‘varimax’ factor-rotation based optimal loadings of participatory variables from FA complemented the two-dimensional PCA biplots with a more detailed understanding of the hidden variance-based grouping patterns. The eigenvalue based factor extraction criteria (eigenvalue >1) resulted in 3 to 4 independent latent factors that maximally explained the total data-system variance among the different study sites (Table 5.4). The FA results are explained based on the magnitudes (i.e., ignoring signs) of the factor loadings given that the transformed data (Z-scores) of the participatory variables consisted of both positive and negative values.

The radiation and heat flux variables (except for FG) had relatively high loadings (0.68 to 0.98) on the first factor (Factor-1) for all but the Mead Irrigated study sites (Table 5.4); NEE also loaded relatively highly (-0.63 to -0.70) on Factor-1 for these sites, reiterating strong carbon flux linkages of the radiation-heat flux variables (i.e., RN, PAR, SHF and LHF). For Mead Irrigated site, the third factor (Factor-3) showed high loadings of NEE (-0.78) and LHF (0.87), as well as moderate loadings of RN (0.55) and PAR (0.50); Factor-1 had a weak loading (-0.16) of NEE and high loadings (0.76 to 0.91) for RN, PAR, SHF and FG. The FG loaded moderate to highly (0.50 to 0.76) on Factor-1 for Niwot Ridge, Sylvania and Fermi Prairie sites. Other notable loadings of FG were apparent on the second factor (Factor-2) for the Everglades (0.48) and Morgan (0.75) sites (Table 5.4).

Table 5.4: Dominant latent factors and optimized loadings of variables for the diverse ecosystem study sites.

Ecosystems	Factor	RN	PAR	LHF	SHF	FG	TA	TS	VPD	SWC	WS	UST	C _{CO2}	NEE
Fermi Prairie	F1	0.95	0.96	0.77	0.80	0.76	0.21	0.05	0.47	-0.07	0.10	0.19	-0.26	-0.70
	F2	0.25	0.21	0.45	-0.15	0.19	0.95	0.94	0.65	-0.29	-0.11	-0.14	-0.21	-0.24
	F3	0.15	0.14	0.13	0.09	0.16	-0.03	-0.08	0.13	0.12	0.90	0.97	-0.30	-0.05
Sylvania Wilderness	F1	0.97	0.96	0.71	0.91	0.65	0.17	0.09	0.44	0.01	-0.01	0.22	-0.30	-0.67
	F2	0.19	0.22	0.35	0.02	0.35	0.98	0.90	0.63	-0.03	-0.06	-0.04	-0.34	-0.07
	F3	0.10	0.08	0.06	0.08	0.11	-0.03	-0.10	0.01	-0.01	0.85	0.97	0.01	-0.01
	F4	0.07	0.06	0.53	-0.14	0.17	0.09	0.16	0.33	-0.13	0.06	0.02	-0.34	-0.37
Niwt Ridge	F1	0.98	0.98	0.68	0.91	0.50	0.17	0.09	0.26	0.10	-0.01	0.11	-0.16	-0.64
	F2	0.16	0.20	0.24	0.07	0.36	0.94	0.77	0.81	0.21	-0.14	-0.15	-0.33	-0.23
	F3	-0.03	0.01	0.12	0.01	-0.03	-0.20	-0.22	-0.05	-0.15	0.95	0.98	-0.05	0.05
Florida Everglades	F1	0.96	0.97	0.71	0.79	0.24	0.35	-0.02	0.68	-0.09	0.23	0.31	-0.13	-0.66
	F2	0.17	0.12	0.31	0.04	0.48	0.93	0.82	0.33	0.15	-0.04	-0.03	-0.33	-0.11
	F3	0.18	0.17	0.26	0.20	0.08	0.01	-0.15	0.18	-0.10	0.83	0.93	-0.16	-0.21
	F4	-0.01	0.04	-0.13	0.43	0.41	0.10	-0.22	0.22	-0.64	0.18	0.18	0.09	-0.25
Mead Irrigated	F1	0.77	0.80	0.30	0.91	0.76	0.13	0.07	0.36		0.12	0.18	-0.13	-0.16
	F2	0.25	0.23	0.32	0.04	0.16	0.97	0.93	0.59		-0.13	0.15	-0.07	-0.16
	F3	0.55	0.50	0.87	-0.01	0.26	0.18	0.15	0.29		-0.01	0.18	-0.39	-0.78
	F4	0.17	0.17	0.13	0.13	0.15	0.01	-0.07	0.18		0.89	0.88	-0.34	0.004
Morgan Forest	F1	0.96	0.96	0.71	0.83	0.28	0.21	0.16	0.42	-0.09	-0.28	0.09	-0.27	-0.63
	F2	0.25	0.27	0.43	-0.08	0.75	0.97	0.87	0.65	-0.35	-0.05	-0.17	-0.22	-0.38
	F3	-0.07	-0.08	-0.14	0.12	0.08	-0.11	-0.24	-0.03	0.18	0.82	0.94	0.06	0.21

Notes: F1-4 refer to four different factors. Bold indicate variables having moderate to high loadings on factors. 'Blank' refers to 'missing data'.

In general, the temperature variables (TA and TS) loaded highly (0.77 to 0.98) on Factor-2, which had low to moderate loadings (-0.07 to -0.38) of NEE. VPD exhibited moderate to high loadings on Factor-1 (0.68) for the Everglades and on Factor-2 (0.59 to 0.81) for the other study sites. SWC showed low to moderate loadings (-0.03 to -0.35) on Factor-2 among the different ecosystem sites. For the Florida Everglades, however, SWC loaded relatively highly (-0.64) on Factor-4 that had a notable loading of NEE (-0.25); suggesting an important linkage between SWC and NEE for the coastal wetland site characterized by a short hydroperiod.

The aerodynamic drivers (WS and UST) loaded heavily on Factor-3 for most ecosystems (0.82 to 0.98) and on Factor-4 for the Mead Irrigated site (0.88 to 0.89). The corresponding low loadings (0.004 to ± 0.21) of NEE on Factor-3 and 4 suggested relatively weak linkages of the aerodynamic variables with NEE at the hourly scale. The ambient C_{CO_2} did not demonstrate strong loadings on any factor; however, the multi-factor associations of C_{CO_2} with low to moderate loadings (0.09 to -0.39) indicated a moderate linkage between ambient CO_2 concentration and NEE.

5.3.4 Climatic and environmental process components driving NEE

Overall, the explanatory PCA and FA identified similar relative linkages and grouping patterns of climatic and ecohydrologic process variables among the six diverse ecosystems that represented a high geographical gradient and range in vegetation and land uses, climate, hydrology, and soil formation. Based on the biplots (Figure 5.3) and factor loadings (Table 5.4), the radiation and heat fluxes (PAR, RN, LHF, SHF and FG) were grouped together as the most dominant process cluster and termed the ‘radiation-energy’ component, which included the strongest drivers (and covariates) of hourly NEE. Although FG sometimes showed the highest loadings outside of Group-A (see Figure 5.3f) or Factor-1 (e.g., Factor-2 for the Everglades and Morgan Forest sites; see Table 5.4), we included FG in the ‘radiation-energy’ component considering the energy

balance closure. Subject to hydrologic process linkages and close associations, the temperature and hydrologic variables (TA, TS, VPD and SWC) were grouped together and termed the ‘temperature-hydrology’ component that represented the moderately strong drivers of hourly NEE. The group of wind speed and friction velocity (WS and UST) was termed the ‘aerodynamic’ component, which represented the relatively weak drivers of NEE. The near-canopy atmospheric C_{CO_2} was termed the ‘ambient atmospheric CO_2 ’ component for quantifying its relative linkage with NEE.

5.3.5 Quantitative relative linking of NEE through explanatory PLSR

Based on a 10-fold cross-validation based optimization, incorporation of the first 3 to 4 PLS components provided the optimal PLSR models for standardized NEE (i.e., Z-scores); resulting in minimum prediction errors (minimum AIC) and explaining the most variance (maximum R^2) of NEE among the different study ecosystems (Figure 5.4). The finding was consistent with the outcomes of PCA and FA. PLSR model calibrations (training) and validations (testing with independent data) with the optimal number of PLS components demonstrated good fitting efficiencies (calibration $R^2 = 0.57-0.77$; validation $R^2 = 0.58-0.75$) and accuracies (calibration RSR = 0.48-0.66; validation RSR = 0.50 to 0.65) (Table 5.5). Model residuals were approximately normally distributed representing constant variances (not shown).

Relative linkages of NEE with the climatic and ecohydrological variables were quantified based on variable importance in the projection (VIP) scores (Figure 5.5) and absolute values of regression coefficients (BETA) (Table 5.5), as obtained from the optimal models upon transformation from the orthogonal (PLS) domain to the original variable domain. Since Z-scores of the participatory variables had both positive and negative values, the sign of BETA was not considered to avoid misleading interpretations on the type (e.g., mutual increase or decrease) of relative NEE linkages of the climatic and ecohydrological variables. The VIP scores

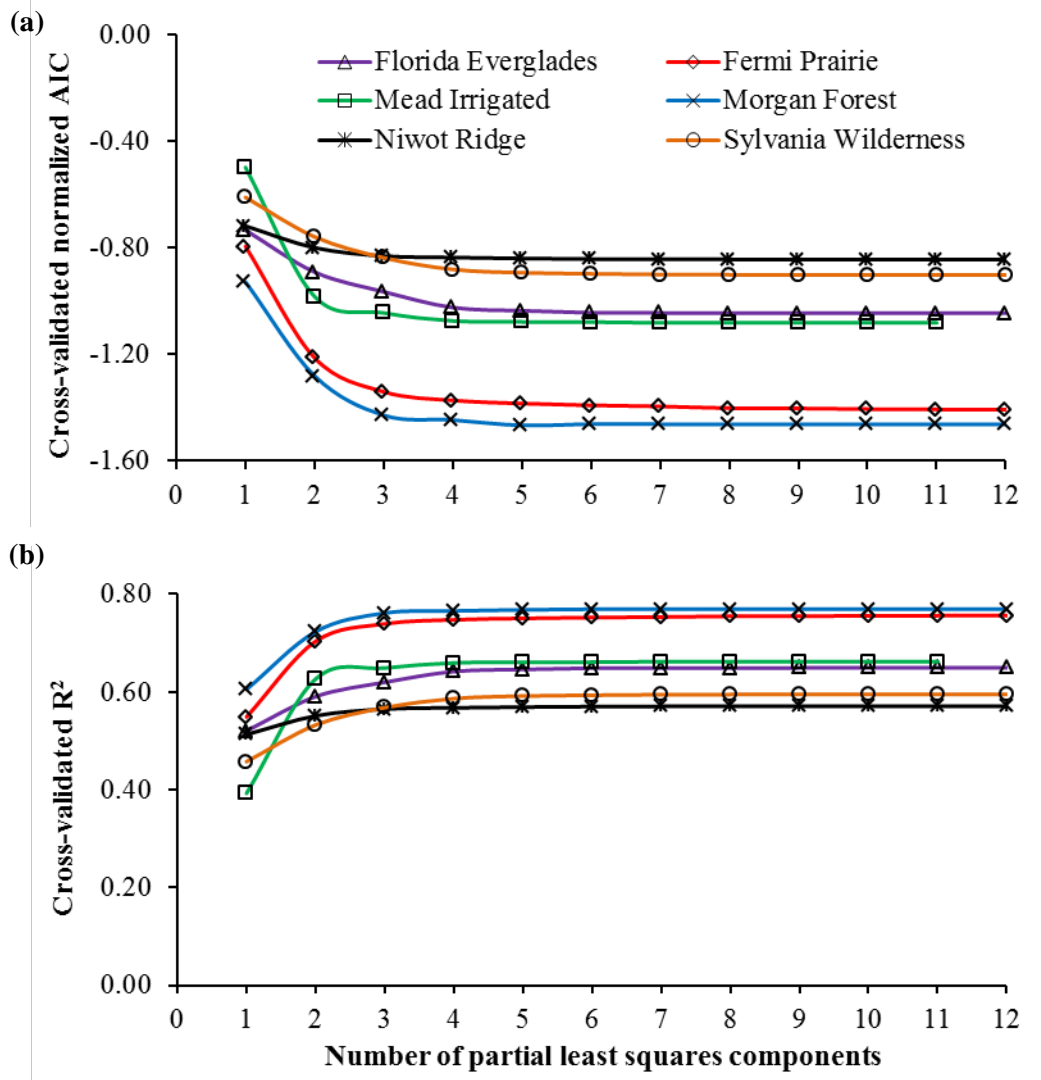


Figure 5.4: Plot of cross-validated (a) normalized AIC and (b) fitting efficiency (R^2) of NEE with the number of partial least squares components.

demonstrated almost similar linkage patterns among different study ecosystems (Figure 5.5).

Variables (LHF, RN, PAR, SHF, FG) representing the ‘radiation-energy’ group had VIP scores greater than or close to unity (1.0) for most ecosystems, suggesting their higher importance and linkages with NEE (see Kuhn and Johnson, 2013). LHF was the most influential predictor (highest VIP scores) of NEE for all but the Everglades site, where SHF appeared to be the strongest predictor. Overall, the ‘temperature-hydrology’ group variables showed relatively

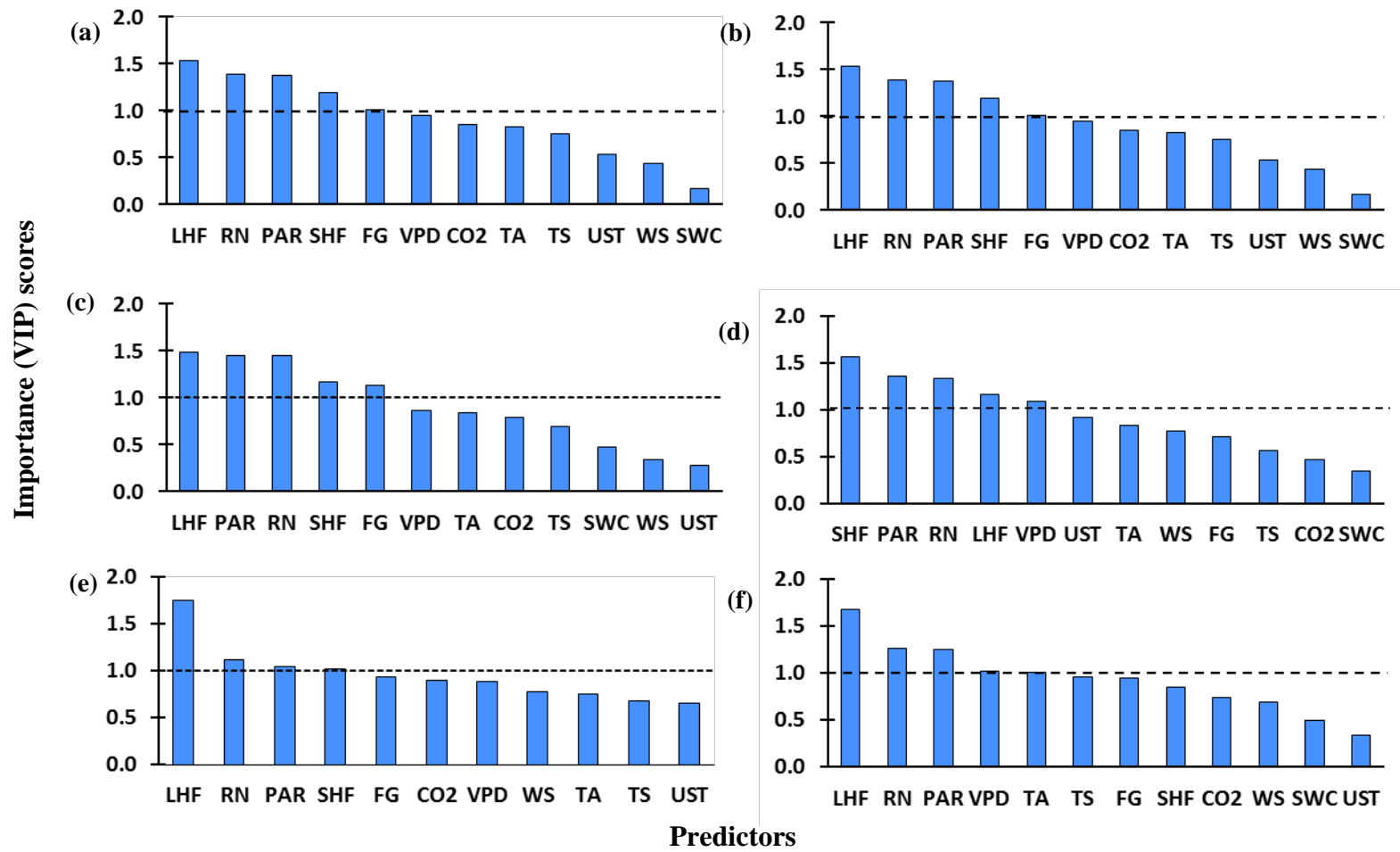


Figure 5.5: Variable importance on the projection (VIP) scores for different predictors of NEE for (a) Fermi Prairie (IL), (b) Sylvania Wilderness (MI), (c) Niwot Ridge Forest (CO), (d) Everglades (FL), (e) Mead Irrigated (NE), and (f) Morgan Monroe State Forest (IN). Dashed line indicates variables with VIP score greater than unity (1.0).

Table 5.5: Coefficients (BETA) of standardized (Z-score) PLSR models of NEE for different study ecosystems.

Variables	Fermi Prairie	Sylvania Wilderness	Niwot Ridge	Florida Everglades	Mead Irrigated	Morgan Forest
RN	-0.25	-0.22	-0.16	-0.06	-0.27	-0.23
PAR	-0.27	-0.21	-0.16	-0.11	-0.20	-0.19
LHF	-0.67	-0.55	-0.28	-0.21	-0.64	-0.56
SHF	0.30	0.09	-0.05	-0.52	0.19	0.20
FG	0.03	0.04	-0.13	0.09	0.20	0.11
TA	0.12	0.12	0.01	-0.14	0.04	-0.04
TS	0.06	0.08	0.06	0.09	0.05	-0.13
VPD	0.11	0.12	0.04	0.16	0.12	0.11
SWC	0.01	0.02	-0.11	0.02		-0.04
WS	0.07	0.00	0.06	0.18	0.15	0.06
UST	0.05	0.05	0.04	-0.18	-0.02	0.02
C _{CO2}	0.11	0.17	0.22	0.16	0.11	0.12
PLS components	4	4	3	4	4	3
<i>Calibrations</i>	<i>2007-10</i>	<i>2002-05</i>	<i>2008-11</i>	<i>2008-11</i>	<i>2008-11</i>	<i>2005-08</i>
R ²	0.75	0.58	0.57	0.66	0.68	0.77
MSE	0.25	0.42	0.43	0.37	0.32	0.23
RSR	0.50	0.64	0.66	0.58	0.57	0.48
<i>Validations</i>	<i>2011</i>	<i>2006</i>	<i>2012</i>	<i>2012</i>	<i>2012</i>	<i>2009</i>
R ²	0.73	0.58	0.64	0.61	0.61	0.75
MSE	0.27	0.42	0.36	0.39	0.39	0.25
RSR	0.52	0.65	0.60	0.62	0.63	0.50

Notes: ‘Blank’ indicates ‘missing data’. MSE refers to mean square error. RSR, the ratio of root-mean-square error to the standard deviation of observations, was calculated as:

$$RSR = \sqrt{\sum_{i=1}^N (NEE_{i,mod} - NEE_{i,obs})^2 / N} / \sigma_{NEE,obs}; \text{ where } NEE_{i,obs} \text{ and } NEE_{i,mod} \text{ are, respectively, the } i^{th} \text{ observed and predicted values of transformed (Z-score) NEE; } \sigma_{NEE,obs} = 1.0 \text{ is the standard deviation of the observed (Z-score) NEE, and } N \text{ is the sample size}$$

moderate VIP scores (with the exceptions of VPD and TA that had a VIP of 1.0 or higher for some sites, while SWC generally exhibiting small VIPs), referring to their moderate linkages with NEE. For both Mead Irrigated and Florida Everglades sites, TA, TS, UST, and WS had similar VIPs, indicating equally moderate NEE linkages of the ‘temperature-hydrology’ and ‘aerodynamic’ components. Otherwise, the ‘aerodynamic’ group variables appeared to be relatively weak predictors of NEE among the different ecosystems. Similar to the outcomes of PCA and FA, the C_{CO2} demonstrated moderate VIP scores, suggesting moderate linkages of ‘ambient atmospheric CO₂’ with NEE for most study ecosystems.

Table 5.6: Aggregated relative linkages of the major biophysical groups with NEE for different ecosystems.

Study Ecosystems	β_{RHF}/β_T	β_{RHF}/β_W	β_{RHF}/β_C	Radiation-energy contribution (%)	Temperature-hydrology contribution (%)	Aerodynamic contribution (%)	Ambient CO ₂ contribution (%)
Fermi Prairie	4.7	9.6	7.5	69	15	7	9
Sylvania Wilderness	3.4	12.7	3.7	61	18	5	16
Niwot Ridge	2.9	5.4	1.8	48	16	9	27
Florida Everglades	2.5	2.3	3.6	47	19	21	13
Mead Irrigated	5.7	5.1	7.0	66	12	13	9
Morgan Forest	3.8	10.7	5.6	65	17	6	12

Notes: The aggregated relative linkages of the ‘radiation-energy’ (β_{RHF}), ‘temperature’ (β_T), ‘aerodynamic’ (β_W), and ‘ambient atmospheric CO₂’ (β_C) components were computed, respectively, as: $\beta_{RHF} = \sqrt{\beta_{RN}^2 + \beta_{PAR}^2 + \beta_{LHF}^2 + \beta_{SHF}^2 + \beta_{FG}^2}$; $\beta_T = \sqrt{\beta_{TA}^2 + \beta_{TS}^2 + \beta_{VPD}^2 + \beta_{SWC}^2}$;

$\beta_W = \sqrt{\beta_{WS}^2 + \beta_{UST}^2}$; and $\beta_C = \sqrt{\beta_{CO2}^2}$.

Magnitudes of the regression coefficients (BETA) exhibited similar linkage patterns of NEE, compared to that of VIPs, with the four variance-based biophysical components and the associated climatic/ecohydrological variables (Table 5.5). Latent heat flux showed the highest linkage with NEE for all ecosystems except for the Florida Everglades (where sensible heat flux had the highest linkage), underlining a strong connection among the ecosystems' carbon, water and energy fluxes. Given that the regression coefficients had both positive and negative signs, the component-wise linkages (aggregated weights) of the 'radiation-energy' (β_{RHF}), 'temperature-hydrology' (β_T), 'aerodynamic' (β_W) and 'ambient atmospheric CO₂' (β_C) groups were computed, respectively, as: $\beta_{RHF} = \sqrt{\beta_{RN}^2 + \beta_{PAR}^2 + \beta_{LHF}^2 + \beta_{SHF}^2 + \beta_{FG}^2}$;

$\beta_T = \sqrt{\beta_{TA}^2 + \beta_{TS}^2 + \beta_{VPD}^2 + \beta_{SWC}^2}$; $\beta_W = \sqrt{\beta_{WS}^2 + \beta_{UST}^2}$; and $\beta_C = \sqrt{\beta_{CO2}^2}$. The relative linkages of the climatic and ecohydrologic process components with NEE (compared to the 'radiation-energy' component) were obtained by calculating the ratios of β_{RHF}/β_T , β_{RHF}/β_W , and β_{RHF}/β_C . The percent contributions of the four process components to the predictions and explanation of NEE were then computed as: $\beta_{RHF} * 100 / (\beta_{RHF} + \beta_T + \beta_W + \beta_C)$;

$\beta_T * 100 / (\beta_{RHF} + \beta_T + \beta_W + \beta_C)$; $\beta_W * 100 / (\beta_{RHF} + \beta_T + \beta_W + \beta_C)$; and

$\beta_C * 100 / (\beta_{RHF} + \beta_T + \beta_W + \beta_C)$.

Among the different study ecosystems, aggregated relative weights (i.e., linkages) of the 'radiation-energy' component were approximately 2.5-6 times higher than that of the 'temperature-hydrology' component; 2-8 times higher than the 'ambient atmospheric CO₂' component; and 2-13 times higher than the 'aerodynamic' component (Table 5.6). Therefore, the 'radiation-energy' contributed by 47-69% to the predictions of NEE, whereas the 'temperature-hydrology', 'ambient atmospheric CO₂' and 'aerodynamic' groups contributed by, respectively,

12-19%, 9-27% and 5-21%. However, NEE linkages (and percent contributions) of the ‘temperature-hydrology’ and ‘aerodynamic’ groups were nearly equal (moderate) in the Florida Everglades and Mead Irrigated ecosystem sites, reiterating similar findings from PCA (see Fig. 3d,e) and VIP scores. On average among all study ecosystems, the ‘radiation-energy’ component was around 4, 5, and 8 times strongly linked with NEE than that of the ‘temperature-hydrology’, ‘ambient atmospheric CO₂’, and ‘aerodynamic’ components, respectively.

5.4 Discussion

5.4.1 On the variable groupings and their comparative linkages

The grouping and interrelation patterns of process variables and their comparative linkages with hourly NEE were essentially similar among the diverse ecosystem sites. The similarity indicated robustness in ecosystem carbon flux process partitions and linkages across a gradient and range of time (2002-12), geographical locations (latitude: 25⁰N to 46⁰N, longitude: 81⁰W to 106⁰W), vegetation and land uses (wetland, grassland, cropland, deciduous, evergreen and mixed forests), forest stand age and canopy heights (2.9-27m), climatic regimes (tropical, temperate, and continental), hydrology, and soil formations (e.g., sand, silt, and clay). The overall findings of the relatively simple data-analytics study are consistent with previous studies. For example, Schmidt et al. (2011) identified similar comparative dominance of process components on daily vertical CO₂ fluxes across diverse ecosystems by applying a complex neural network method. Both studies concluded that the air temperature had a substantially smaller influence on turbulent carbon fluxes than that of the radiation variables; the relatively weak carbon flux linkages of wind and friction velocities were also corroborated by these studies. Further, Ishtiaq and Abdul-Aziz (2015) reported similar grouping, interrelation and relative carbon flux linkage patterns of climate and environmental variable by applying the data-analytics method to half-hourly data of five different years for five U.S. deciduous forests.

Contrary to many previous studies, instead of considering the heat fluxes (LHF, SHF and FG) as ecosystem's response variables alongside NEE, we included them in the predictor data matrix to quantify their relative linkages with NEE within a large set of climate and ecohydrological variables. Melesse and Hanley (2005) also considered heat fluxes as predictors of carbon fluxes in their machine learning based modeling study. This inclusion of heat fluxes acknowledged the mechanistic linkages of ecosystem's carbon, energy and water fluxes. Heat fluxes maintain balance in atmospheric radiation through evapotranspiration (ET) and turbulent energy diffusion (Sellers et al., 1997). Closure of plants' stomata with increasing transpiration, as well as the associated deceleration of LHF to retain water, ultimately affect ecosystem CO₂ fluxes (Heber et al., 1986). Therefore, radiation and heat fluxes work in an integrated system (photosynthesis-respiration feedback) to drive terrestrial carbon, water and energy cycles.

For all study ecosystems, the 'radiation-energy' component (PAR, RN, LHF, SHF and FG) had the strongest link with NEE, supporting findings from previous research (e.g., Morales et al., 2005; Zhang et al., 2005; Jung et al., 2011; Schmidt et al., 2011). LHF showed the highest linkage with NEE for all ecosystems (except for the Florida Everglades site where SHF had the highest linkage), reemphasizing a strong coupling between ecosystems' carbon, water and energy fluxes. A substantially high water availability at the Everglades wetland site (SWC = 68%; see Table 5.2), compared to other ecosystems (SWC = 10-37%), may have resulted in the lower linkage of NEE with the water fluxes (LHF) than that with SHF.

The 'temperature-hydrology' group (TA, TS, VPD and SWC) emerged as the second prominent biophysical component, displaying a moderate linkage with NEE. Mechanistically, temperature-related variables play an important role in ecosystem's carbon, water and energy fluxes. Haworth et al. (2011) reported significant changes in plants' water and energy fluxes due to transformation of stomatal aperture (short-term) and density (long-term) with the variation of

temperature, water availability and light. High VPD can affect fluxes due to stomatal closure (Loescher et al. 2003); a moderate to high linkage of VPD with NEE was apparent for all our study ecosystems. Ryan and Law (2005) reported a strong influence of air temperature (TA) on autotrophic respiration, as well as that of soil temperature (TS) and soil moisture (SWC) on heterotrophic respiration (e.g., high SWC can lead to an anaerobic condition, hindering oxidation processes). Davidson et al. (1998) reported exponentially decreasing trend of soil respiration with increasing drought stress during summer in a temperate mixed forest, although SWC was negatively correlated with respiration during other seasons. However, Bouma et al. (1997) reported a relatively weak correlation between root respiration and SWC.

SWC can also influence plants' photosynthesis by determining the water potential difference between tree leaves and root systems; Desai (2014) reported a significant positive correlation of soil moisture and photosynthesis at different temporal scales (e.g., daily, monthly, yearly) for a temperate mixed forest. However, our analyses with the hourly data generally showed weak linkages between NEE and SWC among the five different ecosystems. Although the PLSR model coefficient (BETA) of SWC was relatively high for the Niwot Ridge site (see Table 5.5), the suggested moderate linkage with NEE was not corroborated by the corresponding PLSR-VIP score (Figure 5.5), factor loading (Table 5.4), PCA biplot (Figure 5.3c) or correlation coefficient (Table 5.3). Similarly, the moderate NEE linkage of SWC, as suggested by factor analysis, for the Everglades site was not corroborated by PLSR or correlation analysis. Future research should investigate whether aggregation of data in longer time-scales (e.g., daily, weekly, monthly, yearly) reveal a more notable linkage of SWC with NEE.

The relatively weak linkages of hourly NEE with the corresponding 'aerodynamic' component (WS and UST) (except for the Everglades and Mead Irrigated sites) are consistent with similar findings from previous studies (e.g., Wilson et al., 2002). Loescher et al. (2003) also

found an overall insignificant relationship between UST and NEE at small temporal resolutions for a tropical wet forest, while reporting a strong linear relationship of UST with nighttime NEE. However, moderately strong influence of ‘aerodynamic’ component for the Everglades and Mead Irrigated sites (see Table 5.6) reflects that NEE is relatively strongly influenced by boundary layer mixing at the coastal wetland (Florida) and cropland (Nebraska) ecosystems.

Identified moderate linkages of NEE with the ‘ambient atmospheric CO₂’ (i.e., C_{CO2}) for different ecosystems can be supported by mechanistic explanations. Elevated atmospheric CO₂ can enhance photosynthesis by stimulating the processes of carboxylation and oxygenation; this primarily results in a faster plant growth and eventually leads to higher litter production and soil carbon storage (Masle 2000). Further, high atmospheric CO₂ results in a maximum utilization of soil water for plant growth and productivity (Schlesinger 1999). Elevated CO₂ can also reduce the rate of plant’s respiration (Bunch, 2001; Pinelli and Loreto, 2003) by impacting mitochondrial enzymes and plant’s chemical composition (Amthor, 1991, 2000; Gonzàlez-Meler et al., 1996a, b) and/or dark CO₂ fixation (Amthor, 1997; Drake et al., 1999). However, Gonzàlez-Meler et al. (2004) reported much uncertainty regarding the relationship between elevated CO₂ and whole-plant respiration rate.

5.4.2 On the effectiveness of the data-analytics method

The systematic data-analytics framework effectively unraveled hidden robust patterns and process components in the climatic, hydrological and ecological data matrices for diverse ecosystems (representing notable spatiotemporal gradients) using a complementary set of data-mining techniques. Achieved mechanistic insights into the dominant process-drivers and their interrelation patterns can guide the development of relatively user-friendly data-driven and process-based models (e.g., by involving a reduced parameter set and simpler structure) to

robustly predict ecosystem carbon fluxes at local to regional scales. In essence, the data-analytics provides an objective empirical framework to analyze ecosystem carbon-water-energy fluxes *a priori* and develop data-based mechanistic understanding, which will ultimately guide and complement process-based modeling of a warranted complexity. More generally, the data-analytics method can be viewed as an important tool to link statistical and process-based modeling approaches—a challenge in ecological modeling research (Larocque et al., 2011).

Although different layers of data-analytics (Figure 5.2) provided complementary and often similar information on the relative linkages of NEE with the climatic and ecohydrological drivers, some occasional disagreements were also observed. For example, FA suggested relatively moderate linkage between NEE and the aerodynamic component (WS and UST) for Morgan Forest (see Table 5.4), whereas a corresponding lower linkage was found based on the PLSR VIP scores (Figure 5.5c) and coefficients (BETA) (Table 5.5, 5.6). However, the suggested moderate carbon flux linkage of aerodynamic component was inferred based on the relatively moderate loading of NEE on Factor 3, which also had very high loadings of WS and UST (see Morgan Forest in Table 5.4). As such, FA provided an indirect quantification of relative linkages by including both response (NEE) and predictors (climatic/ecohydrological variables) in the data matrices. In contrast, PLSR modeling with standardized data (Z-scores) estimated the relative linkages (BETA, VIP) by maximally linking observed data variances of response and predictors. Therefore, the entire data-analytics methodology (combination of correlation coefficient, PCA, FA, and PLSR modeling), rather than any single analysis step, provides a complete representation of the climatic and ecohydrological linkages of NEE.

Subject to the effect of multicollinearity, which is prevalent in ecological and environmental data sets, a conventional multiple regression approach often provides biased estimates of parameters (regression coefficients), resulting in less reliable models. Stepwise

regression modeling (forward addition or backward elimination of predictors) based on statistical significance could partially resolve multicollinearity; however, this can also exclude mechanistically important drivers of the response variable. The first advantage of PLSR modeling with the optimal number of PLS components (i.e., orthogonal predictors) is that the approach mostly resolves multicollinearity existing in predictor data matrices by fitting the model in the orthogonal domain. The second advantage of an optimal PLSR is that transformation of the fitted coefficients associated with predictor PLS components to the original variable (Z-score) domain provides a relatively unbiased estimation of regression weights (i.e., relative NEE linkages) for different climate and ecohydrological variables. The Z-score based PLSR model is, therefore, a convenient and relatively simple approach to quantify comparative carbon flux linkages of different process variables, compared to a more detailed process-based (e.g., Medvigy et al., 2009) or complex data-driven (e.g., Schmidt et al., 2011) model.

Robust estimation of the developed PLSR models was further enhanced by implementing a 10-fold cross-validation technique with large hourly data sets of different 5-year periods spanning over a decade (2002-2012) for six diverse ecosystems. Despite the linear model structure (while the underlying processes can be inherently non-linear), performance of the hourly-scale PLSR models was notably good (training $R^2 = 0.57-0.77$; validation $R^2 = 0.58-0.75$); compared to existing models such as the piecewise linear regression model of Xiao et al. (2008) for daily NEE ($R^2 = 0.53$), and half-hourly semi-mechanistic (non-linear regression) models of Byrne et al. (2005) for GPP (Michaelis-Menten type model; $R^2 = 0.78-0.81$) and respiration (Arrhenius type exponential model; $R^2 = 0.83-0.86$). However, similar to most existing models (linear or non-linear), our linear PLSR models did not show a good skill in predicting extreme positive (source) and negative (sink) fluxes of NEE, which might have resulted in reduced R^2 . Furthermore, modeling of negative and positive NEE fluxes (representing

the domination of GPP and respiration, respectively) with one set of regression coefficients may have reduced R^2 . Significant temporal variability of NEE (Jarvis et al., 1997; Xu and Baldocchi, 2004) also warrants further investigations into the relative linkages with aggregated data of different temporal scales (e.g., daily, weekly, monthly, seasonal, annual) in future research.

5.4.3 On the reliability and uncertainty of empirical data sets

Modeling and analysis with AmeriFlux data can be impacted by measurement errors and uncertainty of eddy fluxes and relevant climate/environmental variables (Williams et al., 2009). Schmidt et al. (2012) performed a comprehensive analysis of errors and uncertainty associated with the AmeriFlux level-2 data, which were used in our study. They reported low relative errors (including random and systematic instrumental errors) for meteorological variables (<2%), CO₂ fluxes (8.2%), and sensible and latent heat fluxes (1.7% to 5.2%); confirming the high quality of AmeriFlux data for research applications. Previous studies (e.g., Moore, 1986; Soegaard et al., 2000; Berger et al., 2001) also suggested the high quality of data measured by using eddy-covariance techniques. We have also avoided artificial biases (e.g., stemming from gap-filling) by using very large sets (N =16457-35592) of un-filled level-2 data among the six different ecosystems. Furthermore, a two-step data filtering procedure was applied on our un-filled data sets to remove hourly data-panels representing gaps for more than two variables in one panel, and to substantially clean the final data sets off unreasonable spikes (outliers). Therefore, our analysis and modeling results were likely to be least impacted by any superfluous measurement errors and uncertainty of AmeriFlux data.

5.5 Conclusions

Relative linkages of NEE with the corresponding climatic and ecohydrological variables were quantified for six diverse AmeriFlux ecosystems by employing a systematic data-analytics

method. The analytics with hourly-scale big data indicated a robust pattern of the biophysical process components and their comparative linkages with NEE across different ecosystems, representing a large gradient in geographical locations, vegetation and land uses, hydro-climatic regimes, forest stand age and canopy heights, and soil formations. Four variance-based biophysical components adequately described the overall system-data variation for all six ecosystems. ‘Radiation-energy’ group demonstrated the strongest linkage with NEE; exhibiting on average (among different ecosystems) 4, 5 and 8 times stronger linkages than that of, respectively, the ‘temperature-hydrology’ group (moderate linkage), ‘ambient atmospheric CO₂’ group (moderate linkage), and ‘aerodynamic’ group (weak linkage). However, a moderate linkage of ‘aerodynamic’ component was found for the Everglades and Mead Irrigated sites, indicating notable influence of boundary layer mixing on NEE at the coastal wetland (Florida) and irrigated cropland (Nebraska) ecosystems. Latent heat flux showed the highest linkage with NEE for all ecosystems except for the Florida Everglades (where sensible heat flux had the highest linkage), underlining a strong connection among the ecosystems’ carbon, water and energy fluxes.

The data-analytics can be used as an objective empirical framework to achieve crucial mechanistic insights into the processes and dynamics of ecosystem carbon-water-energy fluxes *a priori*, guiding process-based modeling of a warranted complexity. Performance of the developed models (training $R^2 = 0.57-0.77$, validation $R^2 = 0.58-0.75$) indicated effectiveness of multivariate analytics to predict NEE fluxes from diverse ecosystems. Specifically, observed similarities in latent patterns and relative linkages showed the scope of formulating parsimonious empirical models to obtain robust predictions of ecosystem carbon fluxes at local to regional scales. Future research should investigate the relative carbon flux linkages at larger spatiotemporal scales across various ecosystems.

Acknowledgements

The research described in this paper was supported by grants from the National Science Foundation (NSF) (NSF CBET Award No. 1336911) and National Oceanic and Atmospheric Administration (NOAA) through the National Estuarine Research Reserve Association (NERRA) (NOAA NERRA Grant No. NA09NOS4190153). The supports are thankfully acknowledged. We also acknowledge the following AmeriFlux sites for publicly sharing their data records: US-Syv, US-IB2, US-Ne1, US-NR1, US-MMS, and US-Esm; funding for the AmeriFlux data resources was provided by the U.S. Department of Energy's Office of Science. The statements, findings, conclusions, and recommendations are those of the authors and do not necessarily reflect the views of NSF or NOAA.

References

- Abdul-Aziz, O. I., Ishtiaq, K. S., 2014. Robust empirical modeling of dissolved oxygen in small rivers and streams: Scaling by a single reference observation. *J. Hydrol.*, 511, 648-657.
- Akaike, H., 1974. A new look at the statistical model identification. *Automatic Control, IEEE Trans. On*, 19(6): 716-723
- Amthor, J. S., 2000. Direct effect of elevated CO₂ on nocturnal in situ leaf respiration in nine temperate deciduous tree species is small. *Tree Physiol.*, 20(2), 139-144.
- Amthor, J. S., 1997. Plant respiratory responses to elevated carbon dioxide partial pressure. *Advances in carbon dioxide effects research*, 61:35-77.
- Amthor, J. S., 1991. Respiration in a future, higher-CO₂ world. *Plant Cell Environ.*, 14(1), 13-20.
- Archibald, S. A., Kirton, A., Merwe, M. R., Scholes, R. J., Williams, C. A., Hanan, N., 2009. Drivers of inter-annual variability in Net Ecosystem Exchange in a semi-arid savanna ecosystem, South Africa. *Biogeosciences*, 6(2), 251-266.
- Baker, I. T., Prihodko, L., Denning, A. S., Goulden, M., Miller, S., Da Rocha, H. R., 2008. Seasonal drought stress in the Amazon: Reconciling models and observations. *Journal of Geophysical Research: Biogeosciences (2005–2012)*, 113(G1).
- Baldocchi, D. D., 2003. Assessing the eddy covariance technique for evaluating carbon dioxide exchange rates of ecosystems: past, present and future. *Global Change Biol.*, 9(4), 479-492.
- Beer, C., Reichstein, M., Tomelleri, E., Ciais, P., Jung, M., Carvalhais, N., ... Papale, D., 2010. Terrestrial gross carbon dioxide uptake: global distribution and covariation with climate. *Science*, 329(5993), 834-838.
- Belward, A. S. (Ed.), 1996. The IGBP-DIS Global 1 Km Land Cover Data Set "DISCover": Proposal and Implementation Plans: Report of the Land Recover Working Group of IGBP-DIS. IGBP-DIS.
- Berger, B. W., Davis, K. J., Yi, C., Bakwin, P. S., Zhao, C. L., 2001. Long-term carbon dioxide fluxes from a very tall tower in a northern forest: Flux measurement methodology. *J. Atmosp. Ocean. Technol.*, 18(4), 529-542.
- Bouma, T. J., Nielsen, K. L., Eissenstat, D. M., Lynch, J. P., 1997. Estimating respiration of roots in soil: interactions with soil CO₂, soil temperature and soil water content. *Plant and Soil*, 195(2), 221-232.
- Braswell, B.H., Sacks, W.J., Linder, E., Schimel, D.S., 2005. Estimating diurnal to annual ecosystem parameters by synthesis of a carbon flux model with eddy covariance net ecosystem exchange observations. *Global Change Biol.*, 11(2): 335-355

- Bunce, J. A., 2001. Effects of Prolonged Darkness on the Sensitivity of Leaf Respiration to Carbon Dioxide Concentration in C₃ and C₄ Species. *Ann. Bot.*, 87(4), 463-468.
- Byrne, K.A., Kiely, G., Leahy, P., 2005. CO₂ fluxes in adjacent new and permanent temperate grasslands. *Agric. For. Meteorol.*, 135(1), 82-92.
- Carrara, A., Kowalski, A. S., Neirynek, J., Janssens, I. A., Yuste, J. C., Ceulemans, R., 2003. Net ecosystem CO₂ exchange of mixed forest in Belgium over 5 years. *Agric. For. Meteorol.*, 119(3), 209-227.
- Chong, I.G., Jun, C.H., 2005. Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intell Laboratory Systems*, 78(1): 103-112
- Clark, K. L., Gholz, H. L., Moncrieff, J. B., Cropley, F., Loescher, H. W., 1999. Environmental controls over net exchanges of carbon dioxide from contrasting Florida ecosystems. *Ecol. Appl.*, 9(3), 936-948.
- Davidson, E., Belk, E., Boone, R. D., 1998. Soil water content and temperature as independent or confounded factors controlling soil respiration in a temperate mixed hardwood forest. *Global Change Biol.*, 4(2), 217-227.
- Davis, S. M., Gaiser, E. E., Loftus, W. F., Huffman, A. E., 2005. Southern marl prairies conceptual ecological model. *Wetlands*, 25(4), 821-831.
- de Jong, S., 1993. SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intell Laboratory Systems*, 18(3): 251-263
- Desai, A. R., 2014. Influence and predictive capacity of climate anomalies on daily to decadal extremes in canopy photosynthesis. *Photosynth. Res.*, 119(1-2), 31-47.
- Desai, A. R., Bolstad, P. V., Cook, B. D., Davis, K. J., Carey, E. V., 2005. Comparing net ecosystem exchange of carbon dioxide between an old-growth and mature forest in the upper Midwest, USA. *Agric. For. Meteorol.*, 128(1), 33-55.
- Desai, A. R., Helliker, B. R., Moorcroft, P. R., Andrews, A. E., Berry, J. A., 2010. Climatic controls of interannual variability in regional carbon fluxes from top-down and bottom-up perspectives. *J. Geophys. Res.: Biogeosci.*, (2005–2012), 115(G2).
- Dragon, K., 2006. Application of factor analysis to study contamination of a semi-confined aquifer (Wielkopolska Buried Valley aquifer, Poland). *J. Hydrol.*, 331(1): 272-279
- Dragoni, D., Schmid, H. P., Wayson, C. A., Potter, H., Grimmond, C. S. B., Randolph, J. C., 2011. Evidence of increased net ecosystem productivity associated with a longer vegetated season in a deciduous forest in south-central Indiana, USA. *Global Change Biol.*, 17(2), 886-897.
- Drake, B. G., Azcon-Bieto, J., Berry, J., Bunce, J., Dijkstra, P., Farrar, J., ... Wullschlegel, S., 1999. Does elevated atmospheric CO₂ concentration inhibit mitochondrial respiration in green plants?. *Plant Cell Environ.*, 22(6), 649-657

- Falge, E., Baldocchi, D., Olson, R., Anthoni, P., Aubinet, M., Bernhofer, C., ... Wofsy, S., 2001. Gap filling strategies for long term energy flux data sets. *Agric. For. Meteorol.*, 107(1), 71-77.
- Falster, D. S., Westoby, M., 2003. Plant height and evolutionary games. *Trends Ecol. Evol.*, 18(7), 337-343.
- Foley, J. A., Prentice, I. C., Ramankutty, N., Levis, S., Pollard, D., Sitch, S., Haxeltine, A., 1996. An integrated biosphere model of land surface processes, terrestrial carbon balance, and vegetation dynamics. *Global Biogeochem. Cycles*, 10(4), 603-628.
- Gilmanov, T. G., Parton, W. J., Ojima, D. S., 1997. Testing the 'CENTURY' ecosystem level model on data sets from eight grassland sites in the former USSR representing a wide climatic/soil gradient. *Ecol. Modell.*, 96(1), 191-210.
- Gonzalez-Meler, M. A., Taneva, L. I. N. A., Trueman, R. J., 2004. Plant respiration and elevated atmospheric CO₂ concentration: cellular responses and global significance. *Ann. Bot.*, 94(5), 647-656.
- Goulden, M. L., Munger, J. W., FAN, S. M., Daube, B. C., Wofsy, S. C., 1996. Measurements of carbon sequestration by long-term eddy covariance: Methods and a critical evaluation of accuracy. *Global Change Biol.*, 2(3), 169-182.
- Gove, J.H., Holinger, D.Y., 2006. Application of a dual unscented Kalman filter for simultaneous state and parameter estimation in problems of surface-atmospheric exchange. *J. Geophys. Res.: Atmos.*, 111: 1-21
- Grant, R. F., Baldocchi, D. D., Ma, S., 2012. Ecological controls on net ecosystem productivity of a seasonally dry annual grassland under current and future climates: Modelling with ecosys. *Agric. For. Meteorol.*, 152, 189-200.
- Haworth, M., Elliott-Kingston, C., McElwain, J.C., 2011. Stomatal control as a driver of plant evolution. *J. Exp. Bot.*, 62(8): 2419-2423
- Heber, U., Neimanis, S., Lange, O.L., 1986. Stomatal aperture, photosynthesis and water fluxes in mesophyll cells as affected by the abscission of leaves. Simultaneous measurements of gas exchange, light scattering and chlorophyll fluorescence. *Planta* 167(4): 554-562
- Hollinger, D. Y., Aber, J., Dail, B., et al., 2004. Spatial and temporal variability in forest-atmosphere CO₂ exchange. *Global Change Biol.*, 10(10), 1689-1706.
- Hui, D., Wan, S., Su, B., Katul, G., Monson, R., Luo, Y., 2004. Gap-filling missing data in eddy covariance measurements using multiple imputation (MI) for annual estimations. *Agric.For. Meteorol.* 121(1): 93-111
- Ishtiaq, K. S., Abdul-Aziz, O. I., 2015. Relative Linkages of Canopy-Level CO₂ Fluxes with the Climatic and Environmental Variables for US Deciduous Forests. *Environ. Manag.*, 55(4), 943-960.

- Jahan, N., Gan, T.Y., 2013. Developing a gross primary production model for coniferous forests of northeastern USA from MODIS data. *Int. J. Appl. Earth Obs. Geoinf.*, 25: 11-20
- Jarvis, P.G., Massheder, J.M., Hale, S.E., Moncrieff, J.B., Rayment, M., Scott, S.L. 1997. Seasonal variation of carbon dioxide, water vapor, and energy exchanges of a boreal black spruce forest. *J. Geophys. Res: Atmos.*, (1984–2012), 102(D24): 28953-28966
- Jolliffe, I.T., 1993. Principal Component analysis: A beginner's guide-II. Pitfalls, myths and extensions. *Weather* 48 (8): 246-253
- Jung, M., Reichstein, M., Margolis, H. A., Cescatti, A., Richardson, A. D., Arain, M. A., ... Williams, C., 2011. Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. *J. Geophys. Res.: Biogeosci.*, (2005–2012), 116(G3).
- Keenan, T. F., Davidson, E., Moffat, A. M., Munger, W., Richardson, A. D., 2012. Using model-data fusion to interpret past trends, and quantify uncertainties in future projections, of terrestrial ecosystem carbon cycling. *Global Change Biol.*, 18(8), 2555-2569.
- Krinner, G., Viovy, N., de Noblet-Ducoudré, N., Ogée, J., Polcher, J., Friedlingstein, P., ... Prentice, I. C., 2005. A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system. *Global Biogeochem. Cycles*, 19(1).
- Kucharik, C. J., Foley, J. A., Delire, C., Fisher, V. A., Coe, M. T., Lenters, J. D., ... Gower, S. T., 2000. Testing the performance of a dynamic global ecosystem model: water balance, carbon balance, and vegetation structure. *Global Biogeochem. Cycles*, 14(3), 795-825.
- Kuhn, M., Johnson, K., 2013. *Applied Predictive Modeling*. Springer, NY
- Larocque, G. R., Bhatti, J. S., Boutin, R., Chertov, O., 2008. Uncertainty analysis in carbon cycle models of forest ecosystems: research needs and development of a theoretical framework to estimate error propagation. *Ecol. Modell.*, 219(3), 400-412.
- Law, B. E., Sun, O. J., Campbell, J., Van Tuyl, S., Thornton, P. E., 2003. Changes in carbon storage and fluxes in a chronosequence of ponderosa pine. *Global Change Biol.*, 9(4), 510-524.
- Law, B. E., Falge, E., Gu, L. V., Baldocchi, D. D., Bakwin, P., Berbigier, P., ... Wofsy, S., 2002. Environmental controls over carbon dioxide and water vapor exchange of terrestrial vegetation. *Agric. For. Meteorol.*, 113(1), 97-120.
- Li, C., Cui, J., Sun, G., & Trettin, C., 2004. Modeling impacts of management on carbon sequestration and trace gas emissions in forested wetland ecosystems. *Environ. Manag.*, 33(1), S176-S186.
- Litvak, M., Miller, S., Wofsy, S. C., Goulden, M., 2003. Effect of stand age on whole ecosystem CO₂ exchange in the Canadian boreal forest. *J. Geophys. Res.: Atmos.*, 108(D3), 8225.

- Liu, C.W., Lin, K.H., Kuo, Y.M., 2003. Science Application of factor analysis in the assessment of groundwater quality in a blackfoot disease area in Taiwan. *Sci. Total Environ.*, 313(1): 77-89
- Lloyd, J., Taylor, J.A., 1994. On the temperature dependence of soil respiration. *Functi, Ecol.*, 8: 315-323
- Loescher, H. W., Oberbauer, S. F., Gholz, H. L., Clark, D. B., 2003. Environmental controls on net ecosystem-level carbon exchange and productivity in a Central American tropical wet forest. *Global Change Biol.*, 9(3), 396-412.
- Mahbub, P., Ayoko, G.A., Goonetilleke, A., Egodawatta, P., Kokot, S., 2010. Impacts of traffic and rainfall characteristics on heavy metals build-up and wash-off from urban roads. *Environ. Sci. Technol.*, 44(23): 8904-8910.
- Mäkelä, A., Pulkkinen, M., Kolari, P., Lagergren, F., Berbigier, P., Lindroth, A., ... Hari, P., 2008. Developing an empirical model of stand GPP with the LUE approach: analysis of eddy covariance data at five contrasting conifer sites in Europe. *Global Change Biol.*, 14(1), 92-108.
- Masle, J., 2000. The effects of elevated CO₂ concentrations on cell division rates, growth patterns, and blade anatomy in young wheat plants are modulated by factors related to leaf position, vernalization, and genotype. *Plant Physiol.*, 122(4), 1399-1416.
- Matamala, R., Jastrow, J. D., Miller, R. M., Garten, C. T., 2008. Temporal changes in C and N stocks of restored prairie: implications for C sequestration strategies. *Ecol. Appl.*, 18(6), 1470-1488.
- Medvigy, D., Wofsy, S. C., Munger, J. W., Hollinger, D. Y., Moorcroft, P. R., 2009. Mechanistic scaling of ecosystem function and dynamics in space and time: Ecosystem Demography model version 2. *J. Geophys. Res.: Biogeosci.*, (2005–2012), 114(G1).
- Melesse, A.M., Hanley, R.S., 2005. Artificial neural network application for multi-ecosystem carbon flux simulation. *Ecol. Modell.*, 189(3): 305-314.
- Menzer, O., Moffat, A. M., Meiring, W., Lasslop, G., Schukat-Talamazzini, E. G., Reichstein, M., 2013. Random errors in carbon and water vapor fluxes assessed with Gaussian Processes. *Agric. For. Meteorol.*, 178, 161-172.
- Monson, R. K., Sparks, J. P., Rosenstiel, T. N., Scott-Denton, L. E., Huxman, T. E., Harley, P. C., ... Hu, J., 2005. Climatic influences on net ecosystem CO₂ exchange during the transition from wintertime carbon source to springtime carbon sink in a high-elevation, subalpine forest. *Oecologia*, 146(1), 130-147.
- Monson, R. K., Turnipseed, A. A., Sparks, J. P., Harley, P. C., Scott-Denton, L. E., Sparks, K., Huxman, T. E., 2002. Carbon sequestration in a high-elevation, subalpine forest. *Global Change Biol.*, 8(5), 459-478.

- Moore, C.J., 1986. Frequency response corrections for eddy covariance systems. *Boundary-Layer Meteorol.*, 37, 17-35.
- Morales, P., Sykes, M.T., Prentice, I.C., 2005. Comparing and evaluating process-based ecosystem model predictions of carbon and water fluxes in major European forest biomes. *Global Change Biol.*, 11(12): 2211-2233
- Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., Veith, T. L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans. Asabe*, 50(3), 885-900.
- Oechel, W. C., Vourlitis, G. L., Verfaillie, J., Crawford, T., Brooks, S., Dumas, E., ...Zulueta, R., 2000. A scaling approach for quantifying the net CO₂ flux of the Kuparuk River Basin, Alaska. *Global Change Biol.*, 6(S1), 160-173.
- Oliphant, A. J., Grimmond, C. S. B., Zutter, H. N., Schmid, H. P., Su, H. B., Scott, S. L., ...Ehman, J., 2004. Heat storage and energy balance fluxes for a temperate deciduous forest. *Agric. For. Meteorol.*, 126(3), 185-201.
- Otieno, D., Lindner, S., Muhr, J., & Boriken, W., 2012. Sensitivity of peatland herbaceous vegetation to vapor pressure deficit influences net ecosystem CO₂ exchange. *Wetlands*, 32(5), 895-905
- Panda, U.C., Sundaray, S.K., Rath, .P., Nayak, B.B., Bhatta, D. 2006. Application of factor and cluster analysis for characterization of river and estuarine water systems—A case study: Mahanadi River (India). *J. Hydrol.*, 331(3): 434-445
- Papale, D. and Valentini, R., 2003., A new assessment of European forests carbon exchanges by eddy fluxes and artificial neural network spatialization. *Global Change Biol.*, 9: 525–535. doi: 10.1046/j.1365-2486.2003.00609.x.
- Peres-Neto, P.R., Jackson, D.A., Somers, K.M., 2003. Giving meaningful interpretation to ordination axes: assessing loading significance in principal component analysis. *Ecology* 84(9): 2347-2363.
- Pinelli, P., Loreto, F., 2003. 12CO₂ emission from different metabolic pathways measured in illuminated and darkened C₃ and C₄ leaves at low, atmospheric and elevated CO₂ concentration. *J. Exp. Bot.*, 54(388), 1761-1769.
- Post, W. M., Pastor, J., 1996. Linkages—an individual-based forest ecosystem model. *Clim. Change*, 34(2), 253-261.
- Reverter, B. R., Sánchez-Cañete, E. P., Resco, V., Serrano-Ortiz, P., Oyonarte, C., Kowalski, A. S., 2010. Analyzing the major drivers of NEE in a Mediterranean alpine shrubland. *Biogeosciences*, 7(9), 2601-2611.
- Richardson, A.D., Braswell, B.H., Hollinger, D.Y. et al., 2006. Comparing simple respiration models for eddy flux and dynamic chamber data. *Agric. For. Meteorol.* 141(2): 219-234.
- Riley, W. J., Still, C. J., Torn, M. S., & Berry, J. A., 2002. A mechanistic model of H₂18O and C₁₈O fluxes between ecosystems and the atmosphere: model description and sensitivity analyses. *Global Biogeochem. Cycles*, 16(4), 42-1.

- Rustad, L. E. J. L., Campbell, J., Marion, G., Norby, R., Mitchell, M., Hartley, A., ..Gurevitch, J., 2001. A meta-analysis of the response of soil respiration, net nitrogen mineralization, and aboveground plant growth to experimental ecosystem warming. *Oecologia*, 126(4), 543-562.
- Ryan, M. G., Law, B. E., 2005. Interpreting, measuring, and modeling soil respiration. *Biogeochemistry*, 73(1), 3-27.
- Schedlbauer, J. L., Munyon, J. W., Oberbauer, S. F., Gaiser, E. E., Starr, G., 2012. Controls on ecosystem carbon dioxide exchange in short-and long-hydroperiod Florida Everglades freshwater marshes. *Wetlands*, 32(5), 801-812.
- Schedlbauer, J. L., Oberbauer, S. F., Starr, G., Jimenez, K. L., 2010. Seasonal differences in the CO₂ exchange of a short-hydroperiod Florida Everglades marsh. *Agric. For. Meteorol.*, 150(7), 994-1006.
- Schlesinger, W. H., 1999. Carbon sequestration in soils. *Science*, 284(5423), 2095.
- Schmidt, A., Hanson, C., Chan, W.S., Law, B.E., 2012. Empirical assessment of uncertainties of meteorological parameters and turbulent fluxes in the AmeriFlux network. *J. Geophys. Res.: Biogeosci.*, (2005–2012) 117(G4).
- Schmidt, A., Hanson, C., Kathilankal, J., Law, B.E., 2011. Classification and assessment of turbulent fluxes above ecosystems in North-America with self-organizing feature map networks. *Agric. For. Meteorol.*, 151(4): 508-520
- Schwalm, C. R., Williams, C. A., Schaefer, K., Anderson, R., Arain, M. A., Baker, I., ... Verma, S. B., 2010. A model-data intercomparison of CO₂ exchange across North America: Results from the North American Carbon Program site synthesis. *J. Geophys. Res.: Biogeosci.*, (2005–2012), 115(G3).
- Sellers, P. J., Dickinson, R. E., Randall, D. A., Betts, A. K., Hall, F. G., Berry, J. A., ... Henderson-Sellers, A., 1997.. Modeling the exchanges of energy, water, and carbon between continents and the atmosphere. *Science*, 275(5299), 502-509.
- Shir, C. C., Bornstein, R. D., 1977. Eddy exchange coefficients in numerical models of the planetary boundary layer. *Boundary-Layer Meteorol.*, 11(2), 171-185.
- Soegaard, H., Nordstroem, C., Friborg, T., Hansen, B. U., Christensen, T. R., Bay, C., 2000. Trace gas exchange in a high-Arctic valley: 3. Integrating and scaling CO₂ fluxes from canopy to landscape using flux data, footprint modeling, and remote sensing. *Global Biogeochem. Cycles*, 14(3), 725-744.
- Stauch, V.J., Jarvis, A.J. 2006. A semi-parametric gap-filling model for eddy covariance CO₂ flux time series data. *Global Change Biol.*, 12 (9): 1707–1716
- Stoy, P. C., Katul, G. G., Siqueira, M. B., Juang, J. Y., McCarthy, H. R., Kim, H. S.,Oren, R., 2005. Variability in net ecosystem exchange from hourly to inter-annual time scales at adjacent pine and hardwood forests: a wavelet analysis. *Tree Physiol.*, 25(7), 887-902.

- Suyker, A. E., Verma, S. B., Burba, G. G., Arkebauer, T. J., 2005. Gross primary production and ecosystem respiration of irrigated maize and irrigated soybean during a growing season. *Agric. For. Meteorol.*, 131(3), 180-190.
- Suyker, A. E., Verma, S. B., Burba, G. G., Arkebauer, T. J., Walters, D. T., Hubbard, K. G., 2004. Growing season carbon dioxide exchange in irrigated and rainfed maize. *Agric. For. Meteorol.*, 124(1), 1-13.
- Tang, X., Wang, Z., Liu, D., Song, K., Jia, M., Dong, Z., ... Liu, X., 2012. Estimating the net ecosystem exchange for the major forests in the northern United States by integrating MODIS and AmeriFlux data. *Agric. For. Meteorol.*, 156, 75-84.
- Tian, H. Q., Chen, G., Liu, M., Zhang, C., Sun, G., Lu, C., ... Chappelka, A., 2010. Model estimates of ecosystem net Primary productivity, evapotranspiration, and water use efficiency in the southern United States during 1895–2007. *Forest Ecol. Manag.*, 259(7), 1311-1327.
- Thornton, P. E., Running, S. W., Hunt, E. R., 2005. Biome-BGC: Terrestrial Ecosystem Process Model, *Version 4.1.1. model product, Oak Ridge Natl. Lab. Distrib. Act. Arch. Cent.*, Oak Ridge, Tenn., doi:[10.3334/ORNLDAAC/805](https://doi.org/10.3334/ORNLDAAC/805). (Available at <http://daac.ornl.gov>).
- Thornton, P. E., Law, B. E., Gholz, H. L., Clark, K. L., Falge, E., Ellsworth, D. S., ... Sparks, J. P., 2002. Modeling and measuring the effects of disturbance history and climate on carbon and water budgets in evergreen needleleaf forests. *Agric. For. Meteorol.*, 113(1), 185-222.
- Verma, S. B., Dobermann, A., Cassman, K. G., Walters, D. T., Knops, J. M., Arkebauer, T. J., ... Walter-Shea, E. A., 2005. Annual carbon dioxide exchange in irrigated and rainfed maize-based agroecosystems. *Agric. For. Meteorol.*, 131(1), 77-96.
- White, A., Cannell, M. G., Friend, A. D., 1999. Climate change impacts on ecosystems and the terrestrial carbon sink: a new assessment. *Global Environ. Change*, 9, S21-S30.
- Williams, M., Richardson, A. D., Reichstein, M., Stoy, P. C., Peylin, P., Verbeeck, H., ... Wang, Y. P., 2009. Improving land surface models with FLUXNET data. *Biogeosciences*, 6(7), 1341-1359.
- Wilson, K., Goldstein, A., Falge, E., Aubinet, M., Baldocchi, D., Berbigier, P., ... Verma, S., 2002. Energy balance closure at FLUXNET sites. *Agric. For. Meteorol.*, 113(1), 223-243.
- Wilson, K. B., Baldocchi, D. D., Hanson, P. J., 2001. Leaf age affects the seasonal pattern of photosynthetic capacity and net ecosystem exchange of carbon in a deciduous forest. *Plant Cell Environ.*, 24(6), 571-583.
- Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intell Laboratory Systems* 58(2): 109-130.

- Wold, S., Johansson, M., Cocchi, M., 1993. PLS-Partial least squares projections to latent structures. In H. Kubinyi (ed) 3D QSAR in Drug design 1: 523-550 Kluwer Academic Publishers, Netherlands.
- Wold, H., 1982. Soft Modelling: The Basic Design and some Extensions. In K.G. Jöreskog & H. Wold (Eds) Systems Under Indirect Observation, Part II. Amsterdam: North Holland Press.
- Wold, H., 1966. Estimation of Principal Components and Related Models by Iterative least squares. In Krishnaiah (ed) Multivariate Analysis: 391-420. Academic Press, New York.
- Wylie, B.K., Fosnight, E.A., Gilmanov, T.G., Frank, A.B., Morgan, J.A, Haferkamp, M.R., Meyers, T.P. 2007. Adaptive data-driven models for estimating carbon fluxes in the Northern Great Plains. *Remote Sens. Environ.*, 106(4): 399-413.
- Xiao, J., Zhuang, Q., Baldocchi, D. D., Law, B. E., Richardson, A. D., Chen, J., ...Torn, M. S., 2008. Estimation of net ecosystem carbon exchange for the conterminous United States by combining MODIS and AmeriFlux data. *Agric. For. Meteorol.*, 148(11), 1827-1847.
- Xu, L, Baldocchi, D.D., 2004. Seasonal variation in carbon dioxide exchange over a Mediterranean annual grassland in California. *Agric. For. Meteorol.*, 123(1): 79-96
- Zhang, Y., Grant, R.F., Flanagan, L.B., Wang, S., Versegny, D.L., 2005. Modelling CO₂ and energy exchanges in a northern semiarid grassland using the carbon-and nitrogen-coupled Canadian Land Surface Scheme (C-CLASS). *Ecol. Modell.*, 181(4): 591-6

Chapter 6: Relative environmental controls and empirical predictive modeling of daytime CO₂ fluxes of coastal salt marshes

Abstract

The relative controls and linkages of different climatic and environmental variables on daytime net CO₂ exchanges of coastal wetlands were determined by using a robust multivariate data-analytics framework. Four different tidal wetlands of Waquoit bay, MA were used as the case study sites, which are subjected to moderate nitrogen gradient. The linkages were then utilized to develop a parsimonious (minimum parameter), nonlinear, data-driven model for the prediction of the CO₂ fluxes of the study wetlands. Dominant controls of radiation (photosynthetically active radiation, PAR), temperature (air and soil temperature), as well as a moderate control of soil salinity on daytime CO₂ exchanges were observed for the salt marshes during the flood condition (when the water table is above the marsh surface). On the contrary, soil salinity, temperature, and water depth demonstrated strong to moderate CO₂ flux linkages during the low tide condition (when the water table is below the marsh surface). A predictive empirical model of the CO₂ fluxes was developed as a nonlinear function of the three most important drivers (PAR, soil temperature, and soil salinity) as determined from the data-analytics framework. The model exhibited very good modeling efficiency (coefficient of determination, $R^2=0.83$) and accuracy (ratio of root-mean-square error to the standard deviation of observations, $RSR=0.41$, and mean absolute error, $MAE=1.40 \mu\text{mole}/\text{m}^2/\text{s}$) in estimating the fluxes. The study facilitates to improve the understanding of the dominant control and mechanism of different environmental drivers in daytime CO₂ exchanges. The predictive model can be used as an ecological engineering tool to aid the development of appropriate GHG offset protocols for restoring and maintaining tidal wetlands.

6.1 Introduction

Coastal salt marshes play a critical role in soil-atmospheric exchanges of one of the major greenhouse gas (GHG) fluxes, carbon dioxide (CO₂), which has a significant global warming potential. Tidal marsh sediments demonstrate the highest ecosystem-specific carbon (C) sequestration rate (Chmura et al., 2003; Duarte et al., 2005; Bridgham et al., 2007; Mcleod et al., 2011). Therefore, tidal wetlands have the potential to play a pivotal role in global warming mitigations. The C storage and sequestration capacity of coastal and marine wetlands are often referred to as “blue carbon” (in contrast to a forest or “green carbon”). "Blue carbon" is an emerging concept for coastal management that can potentially attract much private and public investment in coastal protection and restoration (Nellemann et al., 2009; Crooks et al., 2010, 2011; Morris et al., 2012). However, an overarching science question is how CO₂ flux rates of the coastal wetlands respond to the changes in climate (e.g., temperature, solar radiation), sea level rise (SLR) and inundation (e.g., soil moisture, salinity), and land managements (e.g., nutrient loading). The lack of understanding of wetland C sequestration and C fluxes, and their relative interactions and linkages with the climatic, hydrological, and biogeochemical drivers hinders the development of appropriate models and engineering tools to predict (or project) potential future outcomes. Consequently, the dearth of prediction tools hampers the development of appropriate GHG offset protocols to set guidelines for monitoring requirements for wetlands restoration and maintenance projects. Ultimately, the science and engineering gaps create critical barriers to the management of Blue C stocks in coastal wetland hotspots and their incorporation into a potential C market.

The net CO₂ exchanges of a wetland refer to the net balance of the photosynthesis and respiration (difference between gross primary productivity and both heterotrophic and autotrophic respiration). The daytime net ecosystem exchanges (NEE) are mostly governed by the

photosynthesis (sequestration fluxes). When the fluxes were measured by enclosing the small height (like coastal salt marshes) plants within the chamber, the daytime flux incorporates the daytime respiration from the plants. However, daytime respiration fluxes are significantly lower than the nighttime fluxes for wetlands when all the nighttime flux data are taken into consideration (Juszczak et al., 2012). The daytime NEE or sequestration fluxes are driven by the light as an energy source for photosynthesis. However, ambient climate, hydrology, and biogeochemical processes play an imperative role in regulating the entire process. The comparative contribution of these process variables to the sequestration fluxes vary across wetlands. Therefore, it is important to determine the primary physical controls of the daytime CO₂ flux exchanges of wetlands. Appropriate knowledge on the relative controls of the variables then can guide towards the development of the parsimonious predictive model of wetland CO₂.

Several process-based models are available for the prediction of wetland greenhouse gas (GHG) fluxes (e.g., Cao et al., 1996; Potter 1997; Walter and Heimann 2000; Zhang et al., 2002; van Huissteden et al., 2006). More recent process models of wetland carbon dynamics can include the LPJ-WHy v1.2 (Wania et al., 2009), McGill (St-Hilaire et al., 2010), ecosys (Dimitrov et al., 2010; Grant et al., 2012), peatland DOS-TEM (Fan et al., 2013), WetQual-C (Sharifi et al., 2013), and CoupModel (Jansson, 2012; Metzger et al., 2015). These models are mainly applicable to the freshwater ecosystems but may be adapted for coastal saline water systems. However, mechanistic GHG emissions models are highly detailed, often over parameterized, and requires data for many input variables, while failing to provide spatiotemporally robust predictions. In fact wetland GHG fluxes vary so much in time and space that a robust mechanistic modeling is extremely difficult to achieve with the current state of knowledge. Further, the complexity, expert knowledge requirement, and computational expenses of the mechanistic models hamper their widespread usage as handy ecological engineering tools for ecological and environmental management.

In contrast, relatively simple and parsimonious (involving minimum parameters) data-driven carbon dynamics models can be built on robust empirical patterns hidden in observed data, without necessarily relying on complex scientific hypotheses. Previous research (e.g., Sebacher et al., 1986; Bartlett et al., 1987; Moore and Knowles 1990; Roulet et al., 1992; Dise et al., 1993; Frolking and Crill 1994; Hargreaves and Fowler 1998; Smith et al., 2003; Xing et al., 2005; Yurova et al., 2007; Schedlbauer et al., 2012; Yvon-Durocher et al., 2014) manifested regression-based empirical models mainly for explaining the correlations of GHG fluxes with soil, climate, and other environmental variables. These explanatory empirical models are mostly developed for inland freshwater systems and tailored to be site-specific, and cannot be applied as generic predictive tools for different time and space scales. Generalization in parsimonious parameterization and predictive model development requires explicit mechanistic understanding of relative linkages of the climatic, hydrological and ecological drivers of the wetland CO₂ fluxes. The selection of the best set of unbiased predictive variables for modeling depends on their relative effects on carbon fluxes, co-linearity in a multivariate space, and process descriptions.

Ecological and biogeochemical research has entered into a new regime of data analysis based empirical approaches because of the significant advancements in flux data measurements and computation technologies. Utilization of the 'ecological data' for process understating and robust modeling of GHG fluxes has been a major challenge in the recent time. Application of robust data-analytics methodologies on a multivariate data space could be a viable option to mine latent information of dominant pattern and relative linkages of different climatic, chemical and hydrological variables on CO₂ fluxes for complex coastal wetlands. Ishtiaq and Abdul-Aziz (2015) developed a systematic data-analytics and informatics framework to identify the dominant controls of terrestrial CO₂ fluxes and quantify their relative linkages with various climatic and environmental drivers. Multivariate pattern recognition techniques of principal component and

factor analysis along with Pearson correlation analysis were included in the analytics framework for the similarity based classifications and groupings of the climatic, environmental and ecological variables; interrelation pattern identifications and relative influences. The analysis method also includes explanatory, partial least squares regression (PLSR) models to quantify the relative linkages (weights) of different climatic and environmental stressors on GHG fluxes.

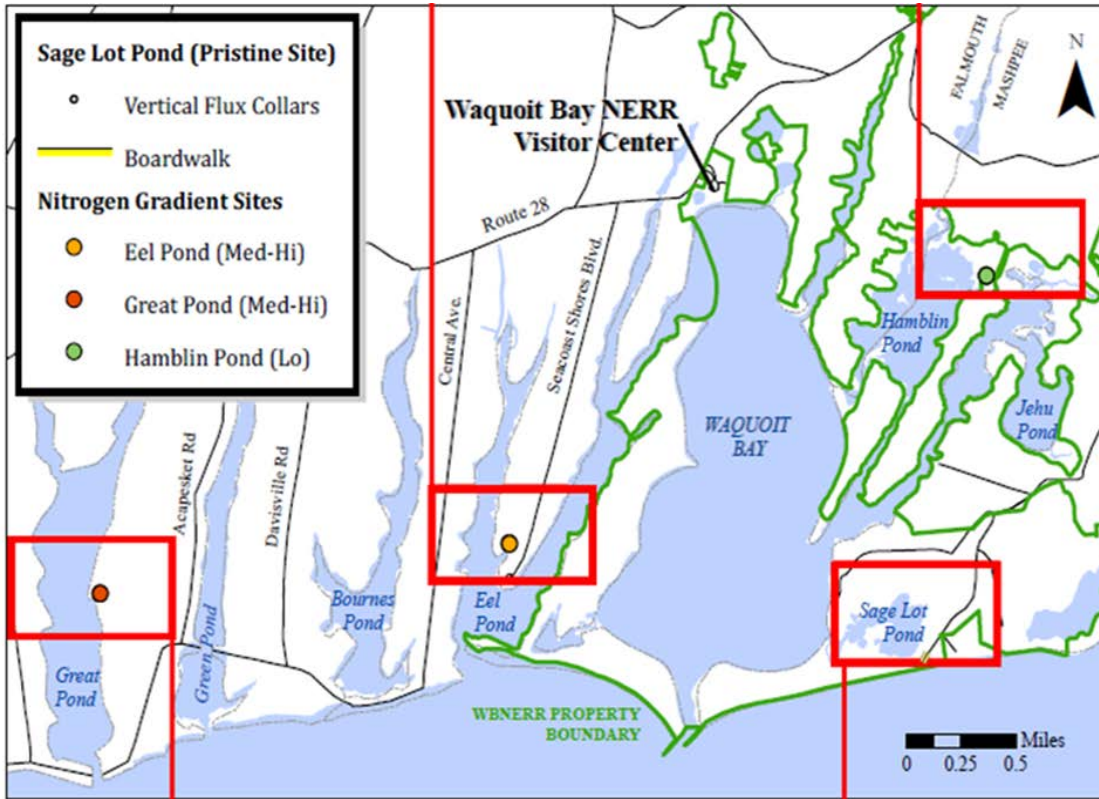
The objectives of this study were to investigate and identify the dominant controls of wetland net CO₂ sequestration fluxes (daytime net ecosystem exchange) and quantify their relative linkages, interrelation patterns and similarity with various hydro-climatic, biogeochemical and environmental drivers by using the systematic multivariate data-analytics and informatics framework developed by Ishtiaq and Abdul-Aziz (2015). The knowledge of similarity and interrelation patterns of the dominant controls were then leveraged to develop a data-driven, parsimonious (minimum parameter) non-linear (e.g., power-law) model for the prediction of wetland CO₂ fluxes.

6.2 Materials and methods

6.2.1 Study sites and data set

The primary data required for the analysis were collected from field measurements at the National Estuarine Research Reserve (NERR) sites of Waquoit Bay, MA (Waquoit Bay website, 2015) as a part of the recently accomplished NOAA-NERRS collaborative project. The Waquoit Bay system comprises of several sub-embayments representing diverse human population densities and an extensive gradient of nitrogen (N) loads per unit area of estuary (as much as 50 fold difference) (Kroeger et al. 2006; Valiela et al. 2000). Four sites representing a range of vegetation, elevation, land use, and inundation were selected for the study: (i) Sage Lot Pond (SL), (ii) Eel Pond (EP), (iii) Great Pond (GP), and (iv) Hamblin Pond (HP) (Figure 6.1). These

selected study ponds have following ecological, hydrological and biogeochemical features: (i) moderate to high level of salinity, (ii) recurrent flooding due to tidal effect, (iii) native *Spartina Alterniflora* (C4 plant) dominated community, and (iv) distinctive low marsh and high marsh



zones (only low marsh zone data was considered for this study)

Figure 6.1: Locations of the four wetland sites in the Waquoit Bay, MA.

Instantaneous daytime CO₂ fluxes and corresponding biogeochemical, hydrological and climatic variables were subsampled from the collected dataset for different days between June to November, 2013 (5 different days in June, 2 days in July, 4 days in August, 2 days in September, 2 days in October, and 1 day in November) (Figure 6.2) in the low marsh zone of the four study ponds. The data collection days represent growing summer and fall seasons (June to November) of coastal salt marshes. Several instantaneous measurements were carried out in each day at different clock hours to cover the entire daytime temporal variability, and both low and high tide

conditions. The complete dataset included net CO₂ sequestration fluxes (can also be termed as daytime net ecosystem exchange, NEE) and seven hydrological, biogeochemical and climatic variables (Table 6.1, Figure 6.1, 6.2): net daytime CO₂ sequestration flux (F_{CO_2} , $\mu\text{mol}/\text{m}^2/\text{s}$), air temperature (AT, °C), soil temperature (ST, °C), photosynthetic active radiation (PAR, $\mu\text{mol}/\text{m}^2/\text{s}$), water depth relative to the marsh surface (WD, m), soil salinity (pore water) (SS, ppt), soil moisture (SM, %) and soil pH (pH). The collected climatic and environmental variables cover potential mechanistic processes (temperature group: AT and ST; radiation group: PAR; hydrology group: WD and SM; and soil biogeochemical group: SS and pH) associated with the CO₂ fluxes.

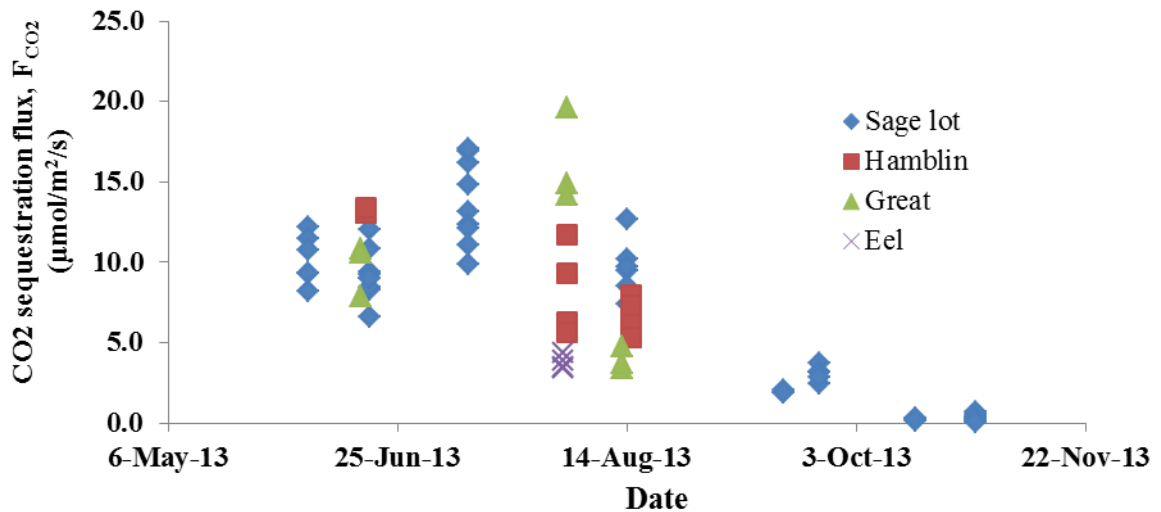


Figure 6.2: Measured CO₂ fluxes for the different days and diurnal hours of 2013 at the four study ponds of Waquoit Bay, MA. Positive values of the fluxes indicate downward (atmosphere to soil) fluxes (i.e., potential sequestration).

The standard GHG flux sign convention, a positive sign represents sequestration was followed for the CO₂ fluxes. Moreover, a positive sign was used when water depth is above the salt marsh surface and –ve sign was used when water depth is below the surface. The CO₂ fluxes were measured by using a Picarro G2501 gas analyzer placed in a chamber (2ft X 2ft X 2ft) in

different adjacent plots of the study sites. The gas concentrations recorded in the gas-analyzer were converted to fluxes through a linear regression from corresponding air temperature by

Table 6.1: Data statistics (mean, standard deviation, maximum, and minimum) of the observed (76 observations) climatic, environmental, and biological variables for the study wetlands.

Variable	PAR ($\mu\text{mol}/\text{m}^2/\text{s}$)	WD (m)	AT ($^{\circ}\text{C}$)	ST ($^{\circ}\text{C}$)	pH	SS (pore water) (ppt)	SM (%)	F _{CO2} ($\mu\text{mol}/\text{m}^2/\text{s}$)
Data	1451.8±531.2	0.09±0.1	27.7±5.9	19.7±4.9	6.8±0.33	29.7±3.8	63.5±3.5	7.5±4.9
Statistics	Max=2080.4 Min=303.7	Max=0.33 Min=-0.08	Max=37.6 Min=12.8	Max=27.8 Min=8.9	Max=7.9 Min=6.1	Max=36 Min=20	Max=68.7 Min=51.2	Max=19.6 Min=0.05

Note: PAR, WD, AT, ST, pH, SS, SM, and F_{CO2} refer, respectively, to photosynthetic active radiation, water depth relative to water marsh surface, air temperature, soil temperature, soil pH, soil salinity (pore water), soil moisture, and daytime CO₂ exchanges of the salt marshes. Positive sign of WD means water table above the salt marshes. Positive sign of F_{CO2} refers downward fluxes (i.e., atmosphere to soil). The unit of each variable was given in the parenthesis

incorporating the ideal gas law. pH of the soil was measured by using a surface probe pH meter (ExStick® Instruments, Nashua, NH).

6.2.2 Data preparation and data statistics

Different QA/QC filters were applied prior to the data analysis and modeling to ensure the quality of the flux and corresponding climatic and environmental data. The first data treatment step was based on the linear regression coefficient, r^2 that was used to linearly convert instantaneous GHG concentrations to fluxes. The r^2 of 0.90 was set as a threshold for the CO₂ fluxes; any instantaneous flux observation that was below the threshold limit was removed from the data matrix along with the corresponding observations of climatic variables. The second data treatment step included the removal of the extreme values (i.e., outliers) from the dataset by applying the interquartile range (IQR) criteria (Tukey, 1977). Any instantaneous observation in each variable outside a range obtained from $Q1-1.5*IQR$ and $Q3+1.5*IQR$ (here, $Q1$ = first quartile, $Q3$ =third quartile, and $IQR=Q3-Q1$) were removed from the dataset along with the entire data panel. Thirdly, each variable was plotted with time to find out any irregularity in their pattern (e.g., sudden high or low values that does not have mechanistic sense) and observations that did not follow the regular pattern were removed based on the mechanistic judgment on wetland biogeochemistry and climate. Gaps created in the data matrix from such removal were not filled to avoid artificial 'gap-filling' biases into the analysis. The final dataset contains 76 observations from four study ponds (fifty two observations from the Sage Lot, eleven observations from the Hamblin pond, four from the Eel pond, and nine observations from the Great pond). Figure 6.2 showed the range of data collection days where each day had multiple data points at different daytime hours.

The F_{CO_2} ranged from very low to as high as 19.62 $\mu\text{mol}/\text{m}^2/\text{s}$ (Table 6.1); the spread of the CO₂ fluxes in magnitude suggested the incorporation of different regimes (low and high CO₂

uptake) of CO₂ sequestration fluxes across the selected study wetlands. All the ponds covered the entire F_{CO2} range except the Eel Pond where available CO₂ flux was comparably less (Figure 6.2). The observed variation in soil temperature (ST) replicated the climate variation of Waquoit Bay since ST can be considered as the surrogate of the ambient air temperature. The mean pH was found to be neutral, and that suggest high productivity of the salt marshes and adjacent plant communities. Sediment salinity (pore water) was considerably high over the growing period (SS ranged from 20-36). The measured PAR covered both cloudless (when PAR is high), and cloudy and evening environments (Low PAR) or the study wetlands. The positive value of mean WD was the indicative of persistent anoxic state in the study sites.

6.2.3 Data Analysis and predictive modeling

The biogeochemical activity and soil-atmospheric exchanges of GHG fluxes from the coastal salt marshes largely depend on the oxygen availability in the sediment pores. Soil water level determines this aerobic-anaerobic state of the salt marshes. Since the coastal wetlands are subjected to diurnal flood and ebb tides (high and low tide), the potential linkages between the different climatic variables and F_{CO2} due to the fluctuating water table were investigated for two conditions.

Condition 1: Water above the salt marsh surface (WD>0)

Condition 2: Water below the salt marsh surface (WD<0).

The final dataset was divided based on the above mentioned conditions, and all the multivariate statistical analysis methodology described below was performed discretely for these two settings. Out of the 76 observation panels, 60 observations that comprise all the four study sites had water above the salt marsh surface. The rest of the 16 observation panels followed the condition 2 subset (WD<0).

Data-analytics Framework

A systematic data-analytics approach (Figure 6.3) was utilized to identify the dominant controls and drivers of wetland F_{CO_2} and quantify their relative linkages. The Pearson correlation analysis showed the univariate correlations among the participatory variables. The principal component analysis (PCA), a multivariate pattern recognition technique (Peres-Neto et al., 2003), was used to group environmental and ecological variables based on their similarity as drivers, testing their interrelation patterns and relative influences using bi-plots. The factor analysis (FA) (Panda et al., 2006) was performed with the “varimax” orthogonal rotation to optimally describe the system variance with fewer latent factors. The optimal latent factors were extracted based on an initial eigenvalue criterion (eigenvalue > 1.0). The partial least squares regression (PLSR) models (Wold et al., 2001) were developed on orthogonal planes to quantify the relative linkages of F_{CO_2} as a function of the independent variable set (PAR, WD, AT, ST, SS, pH, and SM). The PLSR negates the presence of multicollinearity, if any, in the independent variable data set (Kuhn and Johnson 2013). As a part of the data treatment for the statistical analysis, data of all the participatory variables were first log-transformed (base 10) to replicate the non-linear process interactions among the variables. Secondly, all the data were standardized by calculating their respective Z-scores (normalizing each observation’s deviation from the respective mean by the standard deviation) to bring different units onto a comparable scale. The optimum number of PLS components were selected for the PLSR modeling based on the Akaike Information Criterion (AIC) (Akaike 1974), and coefficient of determination (R^2). The selection of optimum PLS components assured minimum modeling error and statistical stability (higher F-value) of the PLSR model. A 10-fold cross-validation (Kuhn and Johnson 2013) was performed to achieve a robust estimate of the PLSR model parameters. Finally, the regression coefficient (BETA) of the log-transformed and normalized (i.e., dimensionless) PLSR models represented the non-linear relative CO_2 flux linkages of the different climate, environmental and hydrologic variables. The

detail of the data-analytics framework for the investigation of relative linkages was described in Ishtiaq and Abdul-Aziz (2015).

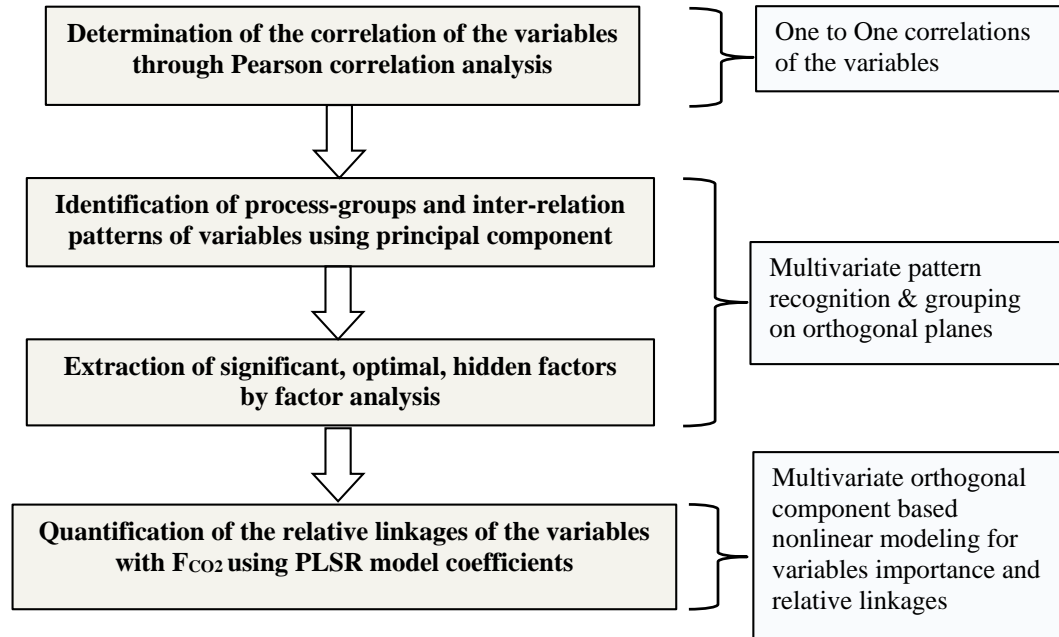


Figure 6.3: The data-analytics methodology to determine the relative linkages of different climatic and environmental variables with F_{CO_2} for the Waquoit Bay salt marshes.

Predictive modeling

The most important and dominant drivers controlling the daytime F_{CO_2} fluxes in the coastal salt marshes were identified based on the data-analytics framework and mechanistic judgment. The selected variables were then utilized to develop a power-law based non-linear parsimonious (i.e., minimum parameter set) model (Equation 1 type) of CO_2 sequestration fluxes for coastal wetlands. The predictive modeling used the original, untransformed (i.e., without log transformation and normalization) data to reduce the prediction uncertainty that may come from different data transformations. Similar to the analysis dataset, the predictive modeling data set incorporated all the observations (76 observations) of the four study ponds to achieve a spatial robustness while fitting the model. Finally, the model was developed by applying a non-linear,

numerical solution based Levenberg-Marquardt least squares algorithm (a modified form of the Gauss-Newton algorithm) (Seber and Wild, 2005, 2003) in the MATLAB environment.

$$F_{CO_2} = 10^a \prod_{i=1} X_i^{b_i} \quad (1)$$

where a and b = model parameters (a is the scaling factor, and b represented shape factors), X = independent variables driving F_{CO_2} .

Because of the absence of independent years of measured data for model validation, the predictive model was developed by using the “leave-group-out cross-validation” (i.e., Monte Carlo cross-validation) approach with 10000 iterations (Kuhn and Johnson 2013). For each iteration, 80% of the data was randomly chosen for model training (i.e. fitting), and the rest of the 20% was used for testing (i.e., validating). The choice of 10000 repetitions reduced the uncertainty of the estimated model parameters since different observations were randomly allocated to the training and testing subsets. Mean values of the parameters obtained from the iterations were used as the final parameter set to predict the F_{CO_2} from the predictor variables for the wetlands.

6.2.4 Measures of modeling performance

The effectiveness of the developed models was assessed by comparing the observed and predicted F_{CO_2} for both explanatory PLSR and predictive models. The coefficient of determination (R^2), ratio of the root-mean-square error to the standard deviations of the observations (RSR), and mean absolute error (MAE) were used as model performance indicators. The R^2 is an indicative of the amount of the observed data variations that is explained by the model. Moffat et al. (2007) provided following four different ranges of R^2 to evaluate a model performance: (i) “very good” ($R^2 > 0.85$), “good” ($0.75 < R^2 \leq 0.85$), “medium” ($0.5 < R^2 \leq 0.75$), and

low ($R^2 \leq 0.50$). The accuracy and error distribution of a model is shown by RSR in dimensionless form. Similar to R^2 , Moriasi et al. (2007) provided a range of RSR values for ranking model accuracy: “perfect to very good” ($0 < \text{RSR} < 0.50$), “good” ($0.50 \leq \text{RSR} < 0.60$), “satisfactory” ($0.60 \leq \text{RSR} < 0.70$), and unsatisfactory ($\text{RSR} \geq 0.70$). The MAE, ranged from zero to infinity, measures the average magnitude and distribution of the individual errors in the dimensional form. The lower the MAE values, the better.

$$R^2 = 1 - \frac{\sum_{i=1}^N (F_{CO_2,obs,i} - F_{CO_2,mod,i})^2}{\sum_{i=1}^N (F_{CO_2,obs,i} - \overline{F_{CO_2,obs}})^2}$$

$$RSR = \frac{\sqrt{\frac{\sum_{i=1}^N (F_{CO_2,obs,i} - F_{CO_2,mod,i})^2}{N}}}{\sigma_{F_{CO_2,obs}}}$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |F_{CO_2,obs,i} - F_{CO_2,mod,i}|$$

Where, N=total number of observations, $F_{CO_2,obs,i}$ and $F_{CO_2,mod,i}$ =the i^{th} observed and predicted CO_2 sequestrating fluxes (respectively), $\overline{F_{CO_2,obs}}$ and $\sigma_{F_{CO_2,obs}}$ =average and standard deviation of all CO_2 flux observations.

6.3 Results

6.3.1 Correlation analysis

Condition 1: Water above the salt marsh surface (WD>0)

Since data was log-transformed and standardized prior to the analysis, only absolute values of the correlations were considered in interpreting the results. The CO_2 sequestration flux, F_{CO_2} was highly correlated ($|r|=0.73-0.94$) with temperature and radiation related variables (AT, ST, and PAR) and moderately correlated ($|r|=0.56$) with soil salinity (SS) (Table 6.2). Relative water

depth (WD) was low to moderately correlated ($|r|=0.27$) with F_{CO_2} in the flooding condition. All the correlations were significant at 95% level of confidence. The radiation and temperature related variables demonstrated significant inter-variable correlations as obtained from the triangular correlation matrix (not shown) implying the presence of a multicollinearity in the independent variable data set.

Condition 2: Water below the salt marsh surface (WD<0)

TA, SS and WD showed moderate non-linear dependence with the F_{CO_2} ($|r|=0.53-63$), while ST and PAR were weakly correlated with the F_{CO_2} (Table 6.2). SM, and pH demonstrated weak to moderate correlation ($|r|=0.30-0.36$) with the CO_2 sequestration fluxes. The presence of the multicollinear structure in the independent variable data set was apparent since moderate correlations were observed between AT-ST, AT-WD, and PAR-pH (not shown).

Table 6.2: Pearson correlation coefficient between the daytime carbon fluxes (F_{CO_2}) and the corresponding climatic and environmental variables for the study area for both WD>0 (water table above the soil marsh) and WD<0 (water table below the soil marsh) conditions.

	PAR	WD	AT	ST	pH	SS	SM
Condition 1 (WD>0)	0.73	-0.27	0.94	0.87	0.18	0.56	0.13
Condition 2 (WD<0)	0.02	-0.55	-0.63	-0.12	0.24	0.53	0.39

Note: PAR, WD, AT, ST, pH, SS, SM, and F_{CO_2} refer, respectively, to photosynthetic active radiation, water depth relative to water marsh surface, air temperature, soil temperature, soil pH, soil salinity, soil moisture, and daytime CO_2 exchanges of the salt marshes.

6.3.2 Relative orientations and groupings of the variables

Condition 1(WD>0)

1st two principal components (PC) explained 64.23% of the total data variances for all the study ponds. The loadings of each variable on the 1st two orthogonal components were presented through biplot (Figure 6.4a). The relative orientation and length of the participatory variables in the biplot revealed the existence of three different variable clusters for the anoxic condition. The

non-orthogonal (0 or 180 degree) orientations between the F_{CO_2} and the group of PAR-AT-ST suggested their higher interrelationship and linkages (all of these variables loaded highly on 1st PC). The relative position of SS and F_{CO_2} in the 2-D plane indicated their moderate linkages. A nearly orthogonal (90 degree) orientation was observed between F_{CO_2} and group of WD-SM-pH (position of SM flipped); suggesting their weak linkages (WD and SM loaded moderately on 2nd PC while F_{CO_2} loaded high on 1st PC). Orthogonal loading between the group of radiation and temperature variables (PAR, AT, and ST) and the group of hydrology variables (WD, SM) suggested the presence of two different process clusters in the data matrix. The radiation-temperature variables were driven by climate and light, and highly linked with the F_{CO_2} . The 2nd cluster (WD and SM) resembled wetland hydrology and found to be nearly uncorrelated with F_{CO_2} . The wetland biogeochemical variables SS showed moderate relation (loaded moderately with both PC's) with F_{CO_2} . pH was weakly linked with the F_{CO_2} since it loaded highly only on 2nd PC.

Condition 2 (WD<0)

The 1st two PC's explained 62.87% of the total data variances (Figure 6.4b). Contrary to the observations of condition 1, no apparent variable based clustering emerged from the biplot for the condition 2. Nearly non-orthogonal orientations between F_{CO_2} and AT, SM, SS and WD suggested their high to moderate linkages. However, relatively small length of SM indicated relatively low dominance. PAR and ST were relatively orthogonally linked with the F_{CO_2} suggesting a weak to moderate CO_2 flux linkage. pH loaded moderately on both of the PCs that suggests moderate F_{CO_2} linkages as F_{CO_2} loaded highly on the 1st PC.

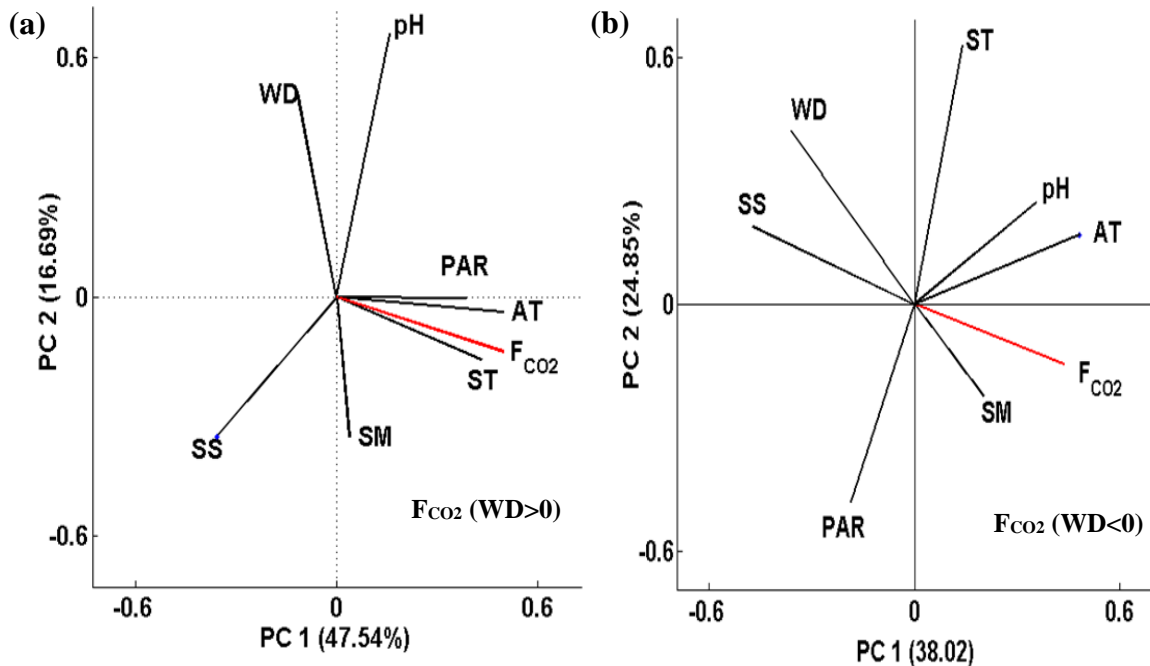


Figure 6.4: Biplots obtained from the loadings of the 1st two principal components (PC) showing the groupings and interrelation patterns of the climatic, environmental, and biological variables for both (a) $WD > 0$ and (b) $WD < 0$ conditions. Percent variance explained by each PC is shown in parenthesis. PAR, WD, AT, ST, pH, SS, SM, and F_{CO_2} refer, respectively, to photosynthetic active radiation, water depth relative to water marsh surface, air temperature, soil temperature, soil pH, soil salinity, soil moisture, and daytime CO_2 exchanges of the salt marshes.

6.3.3 Extracted orthogonal latent factors

Condition 1($WD > 0$)

Three latent factors optimally explained the overall data variances based on the initial Eigenvalue criterion (Eigen value > 1) for the water level above the salt marsh condition (Table 6.3); such optimization while extracting the factors ensured the systematic consideration of most of the data variances (78.8% of the data variances were explained by the 1st three factors) for process description. The factors were ranked on the basis of the percent data explained by each factor; 1st factor explained 46.54%, 2nd factor explained 16.68%, and 3rd factor explained 14.61% of the total data variances. Since, log-transformed Z-scores were used for the analysis, only magnitude of the loadings were considered for the explanation of the FA results.

Temperature variables (AT, ST) loaded highly (0.81-0.98) on 1st factor along with a very high loading (correlation between a factor and a variable) of F_{CO2} (0.84) on this factor; suggesting a strong relationship between them. Since loading of ST was higher than the loading of AT in factor one, the F_{CO2} linkage of ST was higher than the AT. In the second factor, loading of PAR was high (0.92) along with a moderate loading of F_{CO2} (0.45), which indicated an existence of a moderate linkage between them. Although SS loaded highly (-0.93) on factor 3, relatively low loading of F_{CO2} (0.24) on this factor compare to loading of factor 1 and 2 suggested their weak to moderate linkage. The other participatory variables (WD, pH, and SM) did not load significantly on the extracted factors. The lower loading of these variables indicated their very weak linkages and correlation with the F_{CO2} in the flooding condition.

Table 6.3: Extracted latent factors from the factor analysis for the study wetlands for both WD>0 and WD<0 (shaded area in the table) conditions.

	Factors	PAR	WD	AT	ST	pH	SS	SM	F _{CO2}
WD>0	1	0.33	-0.31	0.81	0.98	0.12	-0.29	0.06	0.84
	2	0.92	0.10	0.39	0.06	0.03	-0.21	0.24	0.45
	3	0.20	0.05	0.35	0.10	0.37	-0.93	-0.19	0.24
WD<0	1	0.03	-0.69	0.83	0.12	0.38	-0.77	0.19	0.77
	2	-0.23	0.57	0.37	0.94	0.13	0.13	-0.30	0.06
	3	-0.97	0.07	-0.16	0.31	-0.42	0.05	-0.17	0.06

Note: PAR, WD, AT, ST, pH, SS, SM, and F_{CO2} refer, respectively, to photosynthetic active radiation, water depth relative to water marsh surface, air temperature, soil temperature, soil pH, soil salinity, soil moisture, and daytime CO₂ exchanges of the salt marshes. Bold values indicate most important variables in each factor.

Condition 2 (WD<0)

Similar to the condition 1, three latent factors were extracted for the optimal interpretation of the overall data variances. Three extracted factors explained 76.5% of the variance where the 1st factor explained 38.02%, the 2nd factor explained 24.85%, and the 3rd factor explained 13.62% of the total variations.

AT loaded highly (0.83) and, WD and SS loaded moderately (-0.69 to -0.77) on the 1st factor while loading of F_{CO_2} on this factor is also high (0.7); indicating a strong relationship between AT and F_{CO_2} , and a moderate F_{CO_2} linkages with WD and SS (Table 6.3). ST loaded highly (0.94) on the second factor; low loading of F_{CO_2} on this factor suggested a weak to moderate linkage. Similarly very high loading of PAR (0.97) on the third factor along with a low loading of F_{CO_2} (0.06) on the same factor suggested their weak to moderate linkage. SM and pH did not load significantly on any of the factors indicating a weak F_{CO_2} linkage.

6.3.4 Partial least squares modeling in non-linear domain

The correlation analysis, PCA, and FA facilitated to identify the dominant variables that mostly triggered daytime CO_2 exchanges at varying hydrological conditions. However, these series of statistical investigations did not shed light on the relative magnitude of the linkages between the independent variables and F_{CO_2} . To quantify the linkages, nonlinear partial least squares models were developed for both aerobic and anaerobic conditions with the log-transformed, normalized data to explain the dependent variable (F_{CO_2}) from the independent (PAR, AT, ST, SM, WD, SS, and pH) variables. The AIC optimization (minimum AIC) and R^2 (maximum R^2) led to the selection of 1st two PLS components for both $WD>0$ and $WD<0$ conditions (Figure 6.5) while developing the PLSR models.

Condition 1(WD>0)

In terms of modeling efficiency and accuracy, the PLSR model could be labeled as “very good” and the model is statistically significant ($F_{3,56}=251.02$, $p\text{-value}<0.001$, $R^2=0.95$, $RSR=0.23$, and $MAE=0.18$) (Table 6.4). The observed high modeling efficiency indicated a presence of dominant non-linear liaison between the dependent and independent variables. The estimated model coefficients, BETA were reported in Table 6.4. Temperature variable (AT and ST) and radiation

variable PAR received most of the weights (ST got the height weight) in the regression. Weight of SS and WD were moderate to weak compare to the weight of PAR, AT, and ST.

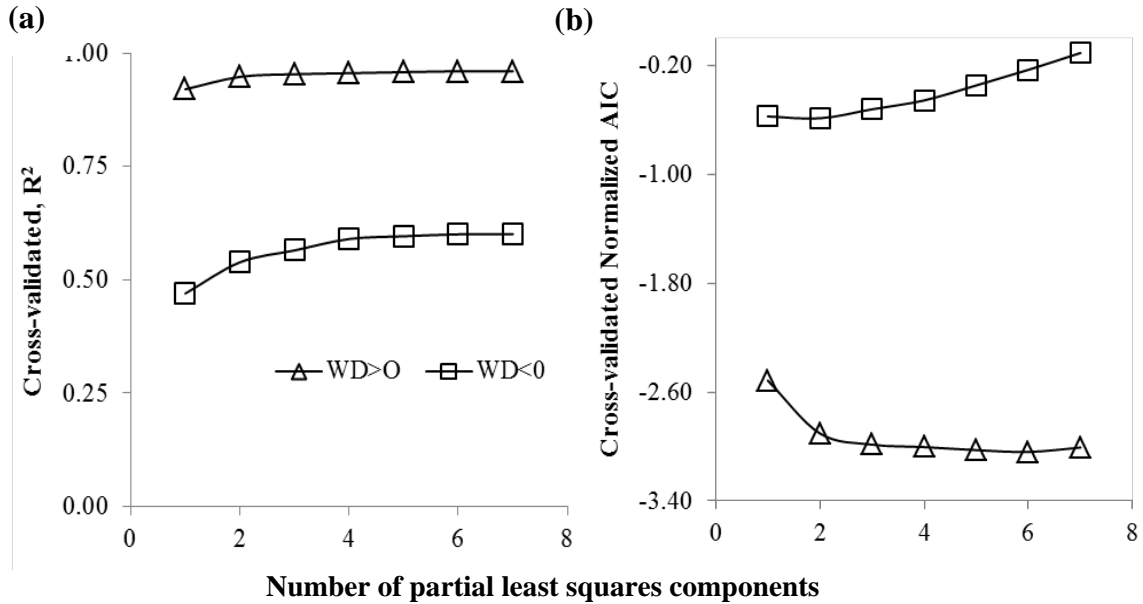


Figure 6.5: Plot of 10-fold cross-validated (a) fitting efficiency, R^2 and (b) Normalized AIC for both $WD>0$ and $WD<0$ conditions of instantaneous F_{CO_2} with the number of incorporated partial least squares (PLS) components instantaneous

Condition 2(WD<0)

The modeling efficiency was not as efficient as compared to condition 1. However, the model was statistically significant based on the overall F-statistics ($F_{3,12}=4.57$, $p\text{-value}<0.05$, $R^2=0.54$, $RSR=0.66$, and $MAE=0.55$) (Table 6.4). Based on the labeling criteria, the model could be ranked as “satisfactory” and “medium”. The lack of ability of the model to describe the system variance efficiently could be linked to the use of relatively small data set (16 observations) while modeling, and complexity in biogeochemical activity beneath the soil when water just recedes due to ebb tide. AT and WD were the most influential in the regression and the Soil salinity (SS) received a moderate weight. ST and PAR were weak to moderately linked with F_{CO_2} based on their BETA value.

Table 6.4: Coefficients (BETA) of the log-transformed, normalized (dimensionless) PLSR models (2 PLS component based) of the daytime F_{CO_2} for both $WD>0$ and $WD<0$ conditions. R^2 , RSR, and MAE refer, respectively, to coefficient of determination, ratio of root-mean-square error to the standard deviation of observations, and mean absolute error.

Variable	PAR	WD	AT	ST	pH	SS	SM	R^2	RSR	MAE
Condition 1 ($WD>0$)	0.28	-0.12	0.34	0.38	-0.02	-0.10	0.09	0.95	0.23	0.18
Condition 2 ($WD<0$)	0.19	-0.34	0.37	0.11	0.03	0.18	0.05	0.54	0.66	0.55

Note: PAR, WD, AT, ST, pH, SS, SM, and F_{CO_2} refer, respectively, to photosynthetic active radiation, water depth relative to water marsh surface, air temperature, soil temperature, soil pH, soil salinity, soil moisture, and daytime CO_2 exchanges of the salt marshes. R^2 , RSR, and MAE refer, respectively, to coefficient of determination, ratio of root-mean-square error to the standard deviation of observations, and mean absolute error.

6.3.5 Predictive model development

Choice of variables for predictive modeling

The 4-layer multivariate statistical provided a clear insight and direction in selecting the most dominant and mechanistically viable drivers for the development of the parsimonious predictive model. Data-analytics did not indicate any firm and direct role of water depth on F_{CO_2} sequestration (daytime NEE) for the $WD>0$ condition. However, WD was moderately linked with F_{CO_2} for $WD<0$ condition. Temperature and PAR were the most common factors that showed high to moderate contribution on F_{CO_2} . Hydrology mediated biogeochemical processes (SS) were moderately linked with F_{CO_2} . As such, PAR, ST, and SS were considered as predictor variables for model development. WD was not considered in the predictor variable set for three reasons: (i) It did not show evidence of a significant linkage for the completely submerged condition, (ii) since the model follows a power law function (Equation 1), inclusion of WD in the predictor variable set lead to the separation of the data set based on the position of the water table related to the marsh surface. If WD showed moderate to strong linkage for both of the cases under consideration, it would be worth to develop two different models for $WD>0$, and $WD<0$ respectively, and (iii) the explanatory non-linear PLSR model only explained 54% of the

variances for the $WD > 0$ condition, which suggests developing a separate model for $WD < 0$ would have indicated similar modeling efficiencies. The predictive model did not use long-transformed and normalized data to model F_{CO_2} ; rather original units of all the participatory variables (ST, PAR, SS, and F_{CO_2}) as reported in Table 6.1 were used for the predictive modeling. Finally, the predictor variable set consisted of PAR, ST, and SS; leading to the development of four parameter model (one scaling factor, a and three shape factors for PAR, ST, and SS, respectively) (equation 1).

Parameter estimation

The model training and fitting were done based on the Monte-Carlo cross-validations with 10000 iterations. The model showed “good” agreement with the observed data for training (i.e., calibration), and testing (i.e., validation) (mean calibration $R^2=0.84$; mean validation $R^2= 0.82$). The histograms of the parameters plotted with the estimations from 10000 simulations indicated that the parameters were normally distributed (Figure 6.6). The standard deviation of each parameter from the mean was quantified (Table 6.5); the standard deviations indicated the uncertainty of the parameters for the model simulation. The standard deviation varied from 0.06-0.40 for all the parameters. The mean p-value of the parameters indicated the statistical significance of the parameters in the nonlinear regression (mean p-value <0.005) (Table 6.5). Finally, instantaneous daytime F_{CO_2} were predicted as a function of instantaneous PAR, ST, and SS from the mean (mean over 10000 iterations) of the estimated parameters (Table 6.5). The model (Equation 2) was statistically significant and the modeling performance can be labeled as “good” and “very good”, respectively for R^2 and RSR ($F_{3,72}=64.08$, p-value <0.001 , $R^2= 0.83$, RSR=0.41, MAE=1.40 $\mu\text{mole}/\text{m}^2/\text{s}$). Figure 6.7 showed the 1:1 plot of the predictions with the observed data for the study ponds. Apart from the two observations of Sage lot and one

observation of Great Pond, model predictions were very impressive suggesting an existence of spatial robustness of the developed data-driven model.

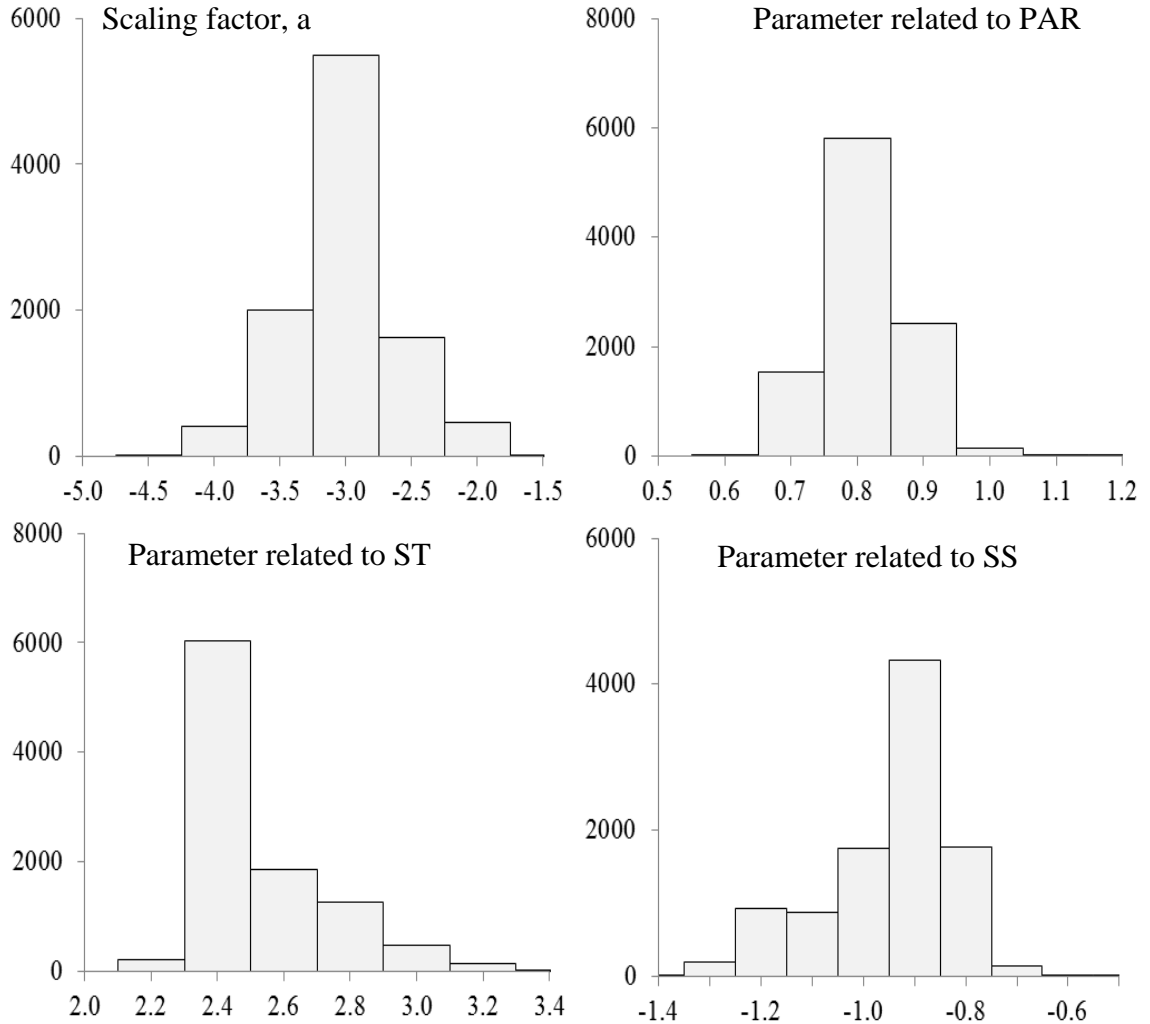


Figure 6.6 Plot shows the histograms of the estimated parameters of the prediction model (option 1: without soil moisture) obtained from the Monte Carlo cross validation through 10000 iterations

$$F_{CO_2} = 10^{-3.26} PAR^{0.76} ST^{2.42} SS^{-1.01} \quad (2)$$

Table 6.5: The estimated model parameters (equation 2) of the predicted model (obtained from the mean of the parameters through 10000 iterations) of daytime F_{CO_2} for the Waquoit bay, MA salt marshes.

	Scaling factor, a	PAR	ST	SS
Mean of estimated parameters	-3.26	0.76	2.42	-1.01
Standard deviation	0.4	0.06	0.19	0.13
P-value	<0.05	<0.05	<0.05	<0.05
R^2		0.83		
RSR		0.41		
MAE ($\mu\text{mol}/\text{m}^2/\text{s}$)		1.40		

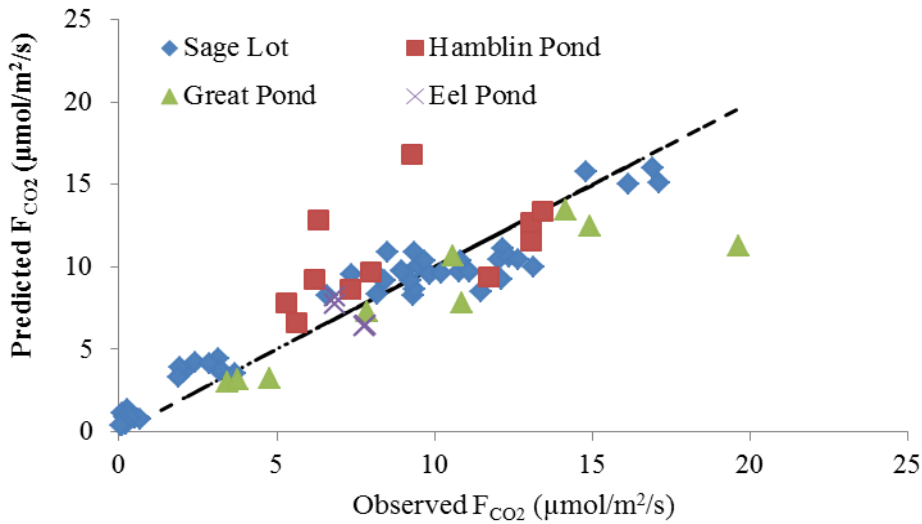


Figure 6.7: Plot of the predicted and observed instantaneous daytime F_{CO_2} of 2013 growing season for the four study ponds of Waquoit Bay, MA. The dashed line represents the 1:1 perfect fit line.

6.4. Discussion

6.4.1 F_{CO_2} linkages of the climatic and environmental variables

The data-analytics revealed high control of the radiation (photosynthetic active radiation, PAR), temperature (air temperature and soil temperature), moderate control of soil salinity, SS (pore

water) on CO₂ exchanges during flood tide condition (when the water table is above the marsh surface). On the contrary, SS, air temperature, and water depth demonstrated strong to moderate CO₂ flux linkages during low tide condition (when the water table is below the marsh surface).

Being a C₄ grass, *Spartina* has the characteristics of high photosynthesis rate (CO₂ fixations) compare to the C₃ plants (*Phragmites*) (Mantlana et al., 2008) because of the absence of photorespiration and significantly less stomatal conductance. Moreover, *Spartina* has the well-developed aerenchyma to adopt with the anoxic and hypoxic conditions imposed by the coastal wetlands (Maricle and Lee 2007). Although a distinct zonation (low marsh and high marsh) was apparent for the New England marshes, this study solely focused on the environment driven response of the low marshes (low elevation zone) on the daytime gas exchanges. A combination of different environmental factors controls the physiological responses of the low marsh *Spartina* under the influence of the tidal cycle.

As apparent from the results (Figure 6.4, Table 6.3, 6.4), light (PAR) showed high F_{CO₂} linkages for WD>0 condition and moderate linkages for WD<0 condition. Overall, fixation of CO₂ by salt marshes largely depends on the availability of sunlight and incident of photons on leaf. Upon diffusion of CO₂ through the stomata opening into the leaf cells, *Spartina* salt marshes follow standard C₄ photosynthesis pathway to generate glucose in the bundle sheath cells. Observed correlation between PAR and temperature variables (ST, AT) suggested their simultaneous contribution on daytime NEE, and the location of the water table (WD>0 and WD<0) determines whether PAR could be the strong driver of F_{CO₂}. The stomatal conductance control the initial CO₂ fixation pathway; stomatal conductance increases with the increase of vapor pressure deficit (vapor pressure deficit increases with temperature) and decrease of xylem water potential (Giurgevich and Dunn 1982). Plants keep the stomata open during high tide (WD>0), and that would probably increase the fixation of light. However, an opposite

phenomenon may occur as water resides below the marsh surface, and stomata resistance starts to develop. The relatively low F_{CO_2} linkages with PAR at $WD < 0$ can also be attributed due to temperature induced daytime heterotrophic respiration. However, Tuittila et al. (2004) reported a decrease of carbon uptake due to the high water table for Sphagnum moss; the study also acknowledged the lack of information and mechanistic judgment on the influence of water level variation in regulating carbon flux exchanges. PAR can also be considered as a surrogate of aboveground biomass; observed higher F_{CO_2} -PAR linkage may also reflect the role of aboveground plant properties in explaining CO_2 uptake.

Temperature, another principal control of CO_2 exchanges, demonstrated very high linkages with F_{CO_2} in both $WD > 0$ and $WD < 0$ conditions. The control of temperature on CO_2 assimilation was also mediated by the ambient CO_2 concentration (Laisk and Edwards, 1997). Ambient temperature controls the activity of the primary C4 photosynthetic enzyme, Rubisco. Rubisco is limited to the low temperature (mostly nighttime) causing decrease in CO_2 assimilation efficiency, whereas at the elevated temperature the turnover rate of Rubisco increases resulting higher CO_2 uptake (Sage and Kubien 2007). Kubien et al. (2003) explained the mechanism of reduced CO_2 exchanges at low temperature as CO_2 leaks out from the bundle sheath. The inability of Spartina to fix CO_2 at low temperature also explained the reason behind the dormancy of the Spartina cord grasses during cold episodes (winter season) at the study sites. Since, the study covers only the daytime F_{CO_2} and study period is limited to the extended growing season, the temperature was adequate to generate high Rubisco turnover. The average air and sediment temperature over the study period was 27.7 and 19.7 °C (Table 6.1), which is higher than the generally acceptable Rubisco activation temperature range (20-25 °C) as reported by Sage and Kubien (2007) from a study on C4 plants. The overall superior controls of PAR, TA, and ST on F_{CO_2} were similar to findings of GUO et al. (2009); the paper investigated the major controls of environmental drivers on tidal wetland carbon balance at short temporal scales. The

study also supported the higher correlation of PAR and F_{CO_2} during inundated condition for *Spartina* salt marshes.

Generally, salt marshes are more tolerant of salt stress; however, productivity of the marshes is limited to the optimum salinity level above which productivity decreases (Lewis and Weber 2002; Wang et al., 2006). The study found low to moderate effect of salinity on F_{CO_2} for $WD > 0$, and higher impact during the low tide ($WD < 0$) condition. The analyzed dataset showed a negative correlation between salinity and F_{CO_2} (Figure 6.8). Soil salinity also could be linked to the root O_2 consumption and stomatal resistance to gas diffusion (Howes et al., 1986). Although no mean difference of salinity was apparent in the data for $WD > 0$ (mean salinity: 29.5 ppt) and $WD < 0$ (salinity: 30 ppt) conditions, higher control of salinity on $WD < 0$ could be explained from the stomatal resistance perspective. The coupled effect of water receding and salinity may cause stomatal resistance to increase and thereby limit the gas exchanges.

The study did not find any notable control of WD on F_{CO_2} during the completely submerged condition; however, a moderate F_{CO_2} linkage was observed for the $WD < 0$ condition. The plant metabolism (aerobic and anaerobic) largely depends on the availability of the diffused O_2 in the roots during water logged condition (Mendelsohn et al., 1981). Sediment oxidation could be linked with the productivity via a positive feedback loop (King et al., 1982; Dacey and Howes 1984). The observed moderate linkages of water depth during the low tide condition may be due to role of WD in regulating CH_4 oxidation in aerobic condition (CH_4 oxidization in the oxidizing layer at the soil surface). The inability of the PLSR model to explain more than 54% of the F_{CO_2} at $WD < 0$ could be due to a complex coupled effect of CO_2 uptake through photosynthesis and CO_2 emission from CH_4 oxidation. The possible higher control of CO_2 emission in this condition could also explain the comparatively lower effect of PAR when water is below the soil surface. As the CO_2 uptake/emission ratio decreases, the relative dominance of PAR decreases

and influence of WD in the exchange mechanism increases. The soil oxygen-reduction potential (Eh) data could also be a good surrogate of the water table to explain the oxygenation and reduction mechanism.

The study revealed minor F_{CO_2} linkages with pH. The pH varied with a narrow range (pH=6.05-7.87) for the entire dataset with an overall mean of 6.86; the mean pH value suggested the overall acidic characteristic and high productivity of the organic salt marshes. Since H^+ is released during redox reaction (i.e., reduction reactions) based microbial activities (Reddy and DeLaune, 2008) pH tends to vary with the O_2 availability in the pores, which is not clearly evident in the analysis. The mechanism of pH in wetland could be more linked to the soil autotrophic respiration (nighttime NEE) than the photosynthesis (daytime NEE).

It was explicitly evident from the analysis that all the factors integrally linked with each other to mechanistically drive the daytime CO_2 exchanges in coastal salt marshes. Although radiation, temperature, and salinity indicated relatively higher control than other participatory variables in the analysis, the role of water availability in the soil pores and also in the xylem tissue in controlling the plant water loss and CO_2 fixation cannot be overemphasized. The described linkages could be altered due to the perturbations caused by the climate change (increased CO_2 concentration, sea level rise). Recent studies suggested higher plant productivity due to elevated CO_2 concentration that increases the marsh surface level to keep pace with the sea level rise (Erickson et al., 2007; Langley et al., 2008). Since photosynthetic rate and stomatal conductance decrease with elevated CO_2 (Nowak et al., 2004; Ainsworth and Long 2005), it could change the quantitative controls of the environmental variables on F_{CO_2} .

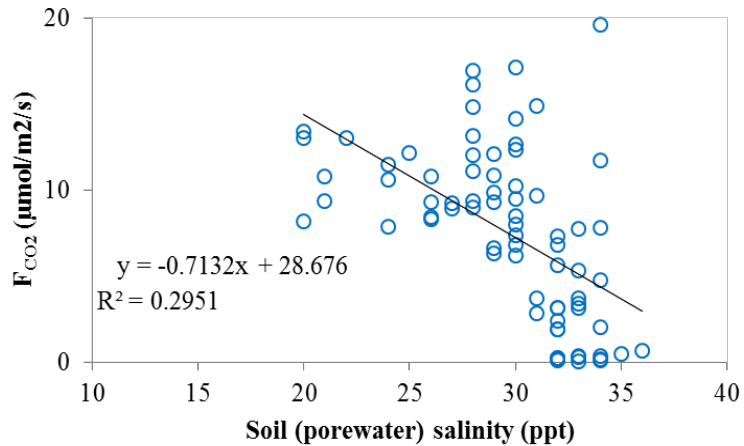


Figure 6.8: Change of daytime CO₂ exchange (F_{CO_2}) with the change of soil salinity for all the four study ponds. The slope of linear fit indicated a negative correlation between them.

6.4.2 Predictive model fitting with a minimum parameter set

Available general photosynthesis models (daytime CO₂ exchanges) mostly incorporated radiation and temperature variables (PAR and temperature) (Monteith 1972; Monteith and Moss 1977; Farquhar et al., 1980) to simulate the photosynthetic biochemistry. Different Q₁₀ based temperature functions (i.e., Arrhenius function) were used to represent the temperature dependency of photosynthesis (Smith and Dukes 2013). Moreover, most of the processes based wetland models that are applicable to the freshwater systems used primary productivity (GPP) to estimate CO₂ related carbon uptake (Cao et al., 1996; Walter and Heimann, 2000). The daytime respiration from the wetland is also a function of temperature. Similarly, radiation and temperature (PAR, ST) were included in the model (Equation 2); however, to account for the saline environment of the coastal salt marshes, based on the results obtained from the data-analytics framework, soil salinity was incorporated in the independent variable set. The inclusion of salinity in the model provided the potential for testing the impact of different climatic event related to the climate change (sea level rise mediated salinity increase) on the daytime NEE. The developed 4-parameter model (one scaling factor, and three shape factors representing the effect of

PAR, ST, and SS, respectively) (Equation 2), therefore, empirically incorporated the primary environmental factors associated with daytime NEE, and the model robustness and parsimony was reserved because of the consideration of such small predictor set. Moreover, the quantitative estimation of the instantaneous daytime fluxes from the small set of environmental variables leveraged the estimation of the long time ecosystem carbon balance (NECB), since CO₂ uptake is one of the important components of NECB.

Since the final model was developed from the mean of the parameters estimated from 10000 iterations and parameters relatively followed the normal distribution with the significantly small standard deviations (Figure 6.6, Table 6.5), the uncertainties of the parameters were less. The final predictive model described 83% (Table 6.5) of the data variances showing the elegance of data-driven empirical modeling in predicting such complex mechanism.

6.4.3 Significance of the estimated parameters

Since the predictive model was developed from the original data set without any data-transformations, the input variables contained their original measuring units (different variable has different unit), thereby it is not feasible to discuss the relative weight of the final estimated parameters regarding magnitude. However, the explanatory PLSR modeling that were performed on the normalized dataset already highlighted the relative importance and interactions between the input variables and the F_{CO₂}. In this section, the significance of the estimated parameters on the basis of their signs were explained.

The value of the scaling parameter, a , was -3.26 (more elaborately $10^{-3.26}=0.00055$) indicating the scaling relationship between the dependent and independent variables. However, a does not represent or explain the underlying shape of the physical parameters in the fitting. Rather a resembles the limit of the equation in the fitting domain. The reason behind the estimation of

the scaling factor as a power of 10 was that it eliminated the model limitation imposed by the scaling factor; if the scaling parameter is estimated without raising to the power of 10 while fitting the data, the value of a can be 0 for a particular combination set during iteration and would then estimate the F_{CO_2} as zero. By raising the a to the power of 10, this limiting factor was avoided.

The shape factors were estimated as 0.76, 2.42, and -1.01 for PAR, ST, and SS, respectively. The positive sign of the parameters in PAR and ST indicated a positive relation with the F_{CO_2} while the negative sign associated with SS indicated a negative relation. However, the observed associations could be described as an independent function to illustrate the behavior of the each parameter in determining the shape of the predictive curve. Since shape factor associated with PAR was between of 0 to 1, the F_{CO_2} will increase with the increase of PAR, but the rate of increase of F_{CO_2} will reduce for a very high value of PAR. In the case of ST, since the shape factor is greater than 1, the F_{CO_2} increases exponentially with the increase of soil temperature. However, beyond a certain air temperature (optimum temperature) the daytime NEE rate (photosynthesis) declines (Sage and Kubien 2007; Smith and Dukes 2013; Atkin and Tjoelker 2003), and there could be a case when the soil temperature and the corresponding air temperature exceeds the thermal optimum. Nevertheless, previous studies reported a typical optimum range of air temperature between 38 °C to 45 °C above which thermal sensitivity rises for C4 plants (Sage 2002; Pittermann and Sage 2000; Sage and Kubien 2007). The maximum air and soil temperature as obtained from the dataset from the study sites were 37.6 and 27.5 oC, respectively, which is below the optimum range. Since the model is fitted with ST and ST remains below AT, the obtained value of the ST parameter did not contradict the existing mechanism.

The CO₂ flux model (Equation 2) was incorporated in an excel spreadsheet along with the wetland CH₄ flux model (see chapter 7) to estimate the net ecosystem carbon balance of a given wetland site. The Excel-based model was named as “BWM Wetland C and GHG Model” and developed as a part of recently completed project “Bringing wetland to market: nitrogen and coastal blue carbon” funded by NOAA. The incorporation of the model into the Excel spreadsheet leverages the prediction of the GHG fluxes in a much easier way. The Excel model can be considered as a user-friendly engineering tool that required minimum input data. The BWM wetland C model is available in the Waquoit Bay National Estuarine Research Reserve website (<http://www.waquoitbayreserve.org/research-monitoring/salt-marsh-carbon-project/>), and detail information of the model and the user-guide is also available as a form of factsheet in the Waquoit Bay website (http://www.waquoitbayreserve.org/wp-content/uploads/BWM-Model-for-Predicting-Greenhouse-Gas-Fluxes_FS_FINAL.pdf)

6.4.4 Generalization of the model

The entire data-analysis and the predictive modeling combined the data from four Waquoit bay wetlands that are subjected to a moderate ecological gradient. The success in model fitting suggested the spatial robustness of the model in the Waquoit bay system. However, to make to model applicable to the regional scale (i.e., New England area) or even more general, the model needs to be validated for new sets of data of different salt marsh sites of New England area along a gradient of nitrogen, salinity, and tidal inundation. The model is only applicable for the low marsh zones; a new model could be developed by incorporating data from the high marsh zone. On the temporal scale, the model was developed by considering the 2013 growing season only. The potential temporal spread of the model could be tested for seasonal and annual upscale values to ensure the applicability of the model in upscaling for the estimation of NECB.

6.5 Conclusions

The relative control and linkages of different climatic and environmental variables on daytime net CO₂ exchanges of coastal salt marshes of four different wetlands of Waquoit Bay, MA were determined by using a robust multivariate data-analytics framework. The linkages were then leveraged to develop a parsimonious (minimum parameter), nonlinear, data-driven model for the prediction of the CO₂ flux exchanges of the study wetlands. The data-analytics revealed high control of the radiation (photosynthetic active radiation, PAR), temperature (air temperature and soil temperature), and moderate control of soil salinity (SS) on CO₂ exchanges during flood tide condition (when water table is above the marsh surface). On the contrary, soil salinity, air temperature and relative water depth demonstrated strong to moderate CO₂ flux linkages during low tide condition (when the water table is below the marsh surface). The empirical predictive model of instantaneous CO₂ fluxes was developed as a nonlinear function of the most three important drivers (PAR, sediment temperature, and SS) as determined from the data-analytics framework. The developed data-driven model could be used as a supplement to the existing process-based models to provide empirical foundation in obtaining critical decisions on “blue carbon” sequestration.

The study facilitates to improve the understanding of the dominant control and mechanism of different environmental drivers in daytime CO₂ exchanges. The developed predictive model could be used to complement the existing process based models. Since the fate of the “blue carbon” under the current change in climate and sea level rise is quite uncertain, the predictive model developed in this study as a component of net ecosystem carbon balance will act as an ecological engineering tool to aid the development of appropriate GHG offset protocols for settling monitoring guideline for tidal wetland restoration and maintenance projects.

References

- Ainsworth, E. A., Long, S. P., 2005. What have we learned from 15 years of free-air CO₂ enrichment (FACE)? A meta-analytic review of the responses of photosynthesis, canopy properties and plant production to rising CO₂. *New Phytologist*, 165(2), 351-372.
- Akaike, H., 1974. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6), 716-723.
- Atkin, O. K., Tjoelker, M. G., 2003. Thermal acclimation and the dynamic response of plant respiration to temperature. *Trends in plant science*, 8(7), 343-351.
- Bartlett, K.B., Bartlett, D.S., Harriss, R.C., Sebacher, D.I., 1987. Methane emissions along a salt-marsh salinity gradient. *Biogeochemistry*, 4(3), 183-202.
- Bridgham, S. D., J. P. Megonigal, J. K. Keller, N. B. Bliss, and C. Trettin. 2007. Wetlands. In: A.W. King, L. Dilling, G.P. Zimmerman, et al. (Eds.), *The First State of the Carbon Cycle Report (SOCCR): The North American Carbon Budget and Implications for the Global Carbon Cycle. A Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research. National Oceanic and Atmospheric Administration, National Climatic Data Center, Asheville, NC, USA*, pp. 139- 148.
- Cao, M., Marshall, S., Gregson, K., 1996. Global carbon exchange and methane emissions from natural wetlands: Application of a processes based model. *Journal of Geophysical Research*, 101(D9), 14399–14414.
- Chmura, G. L., Anisfeld, S. C., Cahoon, D. R., Lynch, J. C., 2003. Global carbon sequestration in tidal, saline wetland soils. *Global biogeochemical cycles*, 17(4).
- Crooks, S., Emmett-Mattox, S., Findsen, J., 2010. Findings of the National Blue Ribbon Panel on the development of a greenhouse gas offset protocol for tidal wetlands restoration and management: action plan to guide protocol development. *Restore America's Estuaries*, Philip Williams & Associates, Ltd., and Science Applications International Corporation.
- Crooks, S., Herr, D., Tamelander, J., Laffoley, D., Vandever, J., 2011. Mitigating climate change through restoration and management of coastal wetlands and near-shore marine ecosystems: challenges and opportunities.
- Dacey, J. W., Howes, B. L., 1984. Water uptake by roots controls water table movement and sediment oxidation in short *Spartina* marsh. *Science*, 224(4648), 487-489.
- Dimitrov, D. D., Grant, R. F., Lafleur, P. M., Humphreys, E. R., 2010. Modeling the effects of hydrology on ecosystem respiration at Mer Bleue bog. *Journal of Geophysical Research: Biogeosciences* (2005–2012), 115(G4).
- Dise, N. B., Gorham, E., Verry, E.S., 1993. Environmental Factors Controlling Methane Emissions from Peatlands in Northern Minnesota. *Journal of Geophysical Research*, 98(D6), 10583-10594, doi:10.1029/93JD00160.

- Duarte, C. M., Middelburg, J. J., Caraco, N. F., 2005. Major role of marine vegetation on the oceanic carbon cycle. *Biogeosciences*, 2(1), 1-8.
- Erickson, J. E., Megonigal, J. P., Peresta, G., Drake, B. G., 2007. Salinity and sea level mediate elevated CO₂ effects on C3–C4 plant interactions and tissue nitrogen in a Chesapeake Bay tidal wetland. *Global Change Biology*, 13(1), 202-215.
- Fan, Z., David McGuire, A., Turetsky, M. R., Harden, J. W., Michael Waddington, J., Kane, E. S., 2013. The response of soil organic carbon of a rich fen peatland in interior Alaska to projected climate change. *Global change biology*, 19(2), 604-620.
- Farquhar, G. D., von Caemmerer, S. V., Berry, J. A., 1980. A biochemical model of photosynthetic CO₂ assimilation in leaves of C3 species. *Planta*, 149(1), 78-90.
- Frolking, S., Crill, P., 1994. Climate controls on temporal variability of methane flux from a poor fen in southeastern New Hampshire: Measurement and modeling. *Global Biogeochemical Cycles*, 8, 385–397
- Grant, R. F., Desai, A. R., Sulman, B. N., 2012. Modelling contrasting responses of wetland productivity to changes in water table depth, *Biogeosciences*, 9, 4215-4231, doi:10.5194/bg-9-4215-2012, 2012.
- Guo, H., Noormets, A., Zhao, B., Chen, J., Sun, G., Gu, Y., Chen, J., 2009. Tidal effects on net ecosystem exchange of carbon in an estuarine wetland. *Agricultural and Forest Meteorology*, 149(11), 1820-1828.
- Hargreaves, K. J., Fowler, D., 1998. Quantifying the effects of water table and soil temperature on the emission of methane from peat wetland at the field scale. *Atmospheric Environment*, 32(19), 3275-3282.
- Howes, B. L., Dacey, J. W. H., Goehring, D. D., 1986. Factors controlling the growth form of *Spartina alterniflora*: feedbacks between above-ground production, sediment oxidation, nitrogen and salinity. *The Journal of Ecology*, 881-898.
- Ishtiaq, K.S., Abdul-Aziz, O.I., 2015. Relative linkages of canopy-level CO₂ fluxes with the climatic and environmental variables for US deciduous forests. *Environmental Management*, 55 (4), 943-960.
- Jansson, P. E., 2012. CoupModel: model use, calibration, and validation. *Transactions of the ASABE*, 55(4), 1335-1344.
- Juszczak, R., Acosta, M., Olejnik, J., 2012. Comparison of daytime and nighttime ecosystem respiration measured by the closed chamber technique on a temperate mire in Poland. *Polish Journal of Environmental Studies*, 21(3), 643-658.
- King, G. M., Klug, M. J., Wiegert, R. G., Chalmers, A. G., 1982. Relation of soil water movement and sulfide concentration to *Spartina alterniflora* production in a Georgia salt marsh. *Science*, 218(4567), 61-63.

- Kubien, D. S., von Caemmerer, S., Furbank, R. T., Sage, R. F., 2003. C4 photosynthesis at low temperature. A study using transgenic plants with reduced amounts of Rubisco. *Plant Physiology*, 132(3), 1577-1585.
- Kuhn, M., Johnson, K., 2013. *Applied predictive modeling*, New York: Springer.
- Kroeger, K. D., Cole, M. L., York, J. K., Valiela, I., 2006. Nitrogen loads to estuaries from waste water plumes: Modeling and isotopic approaches. *Groundwater*, 44(2), 188-200.
- Langley, J. A., McKee, K. L., Cahoon, D. R., Cherry, J. A., Megonigal, J. P., 2009. Elevated CO₂ stimulates marsh elevation gain, counterbalancing sea-level rise. *Proceedings of the National Academy of Sciences*, 106(15), 6182-6186.
- Lewis, M. A., Weber, D. E., 2002. Effects of substrate salinity on early seedling survival and growth of *Scirpus robustus* Pursh and *Spartina alterniflora* Loisel. *Ecotoxicology*, 11(1), 19-26.
- Mantlana, K. B., Arneith, A., Veenendaal, E. M., Wohland, P., Wolski, P., Kolle, O., et al., 2008. Photosynthetic properties of C4 plants growing in an African savanna/wetland mosaic. *Journal of experimental botany*, 59(14), 3941-3952.
- Maricle, B. R., Lee, R. W., 2007. Root respiration and oxygen flux in salt marsh grasses from different elevational zones. *Marine Biology*, 151(2), 413-423.
- Mcleod, E., Chmura, G. L., Bouillon, S., Salm, R., Björk, M., Duarte, C. M., et al., 2011. A blueprint for blue carbon: toward an improved understanding of the role of vegetated coastal habitats in sequestering CO₂. *Frontiers in Ecology and the Environment*, 9(10), 552-560.
- Mendelssohn, I. A., McKee, K. L., Patrick, W. H., 1981. Oxygen deficiency in *Spartina alterniflora* roots: metabolic adaptation to anoxia. *Science*, 214 (4519), 439-441.
- Metzger, C., Jansson, P. E., Lohila, A., Aurela, M., Eickenscheidt, T., Beilelli-Marchesini, L., et al., 2015. CO₂ fluxes and ecosystem dynamics at five European treeless peatlands—merging data and process oriented modeling. *Biogeosciences*, 12(1), 125-146.
- Moffat, A. M., Papale, D., Reichstein, M., Hollinger, D. Y., Richardson, A. D., Barr, A. G., et al., (2007). Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. *Agricultural and Forest Meteorology*, 147(3), 209-232.
- Monteith, J. L., 1972. Solar radiation and productivity in tropical ecosystems. *Journal of applied ecology*, 747-766.
- Monteith, J. L., Moss, C. J., 1977. Climate and the efficiency of crop production in Britain [and discussion]. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 281(980), 277-294.
- Moore, T.R., Knowles, R., 1990. Methane emissions from fen, bog and swamp peatlands in Quebec. *Biogeochemistry*, 11(1), 45-61.

- Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., Veith, T. L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans. Asabe*, 50(3), 885-900.
- Morris, J. T., Edwards, J., Crooks, S., Reyes, E., 2012. Assessment of carbon sequestration potential in coastal wetlands. In *Recarbonization of the Biosphere* (pp. 517-531). Springer Netherlands.
- Nellemann, C., Corcoran, E., (Eds.) 2009. *Blue carbon: the role of healthy oceans in binding carbon: a rapid response assessment*. UNEP/Earthprint.
- Nowak, R. S., Ellsworth, D. S., Smith, S. D., 2004. Functional responses of plants to elevated atmospheric CO₂—do photosynthetic and productivity data from FACE experiments support early predictions?. *New Phytologist*, 162(2), 253-280.
- Panda, U. C., Sundaray, S. K., Rath, P., Nayak, B. B., Bhatta, D., 2006. Application of factor and cluster analysis for characterization of river and estuarine water systems—A case study: Mahanadi River (India). *Journal of Hydrology*, 331(3), 434-445.
- Peres-Neto, P.R., Jackson, D.A., Somers, K.M., 2003. Giving meaningful interpretation to ordination axes: assessing loading significance in principal component analysis. *Ecology*, 84(9), 2347-2363.
- Pittermann, J., Sage, R. F., 2000. Photosynthetic performance at low temperature of *Bouteloua gracilis* Lag., a high-altitude C₄ grass from the Rocky Mountains, USA. *Plant, Cell & Environment*, 23(8), 811-823.
- Potter, C. S., 1997. An ecosystem simulation model for methane production and emission from wetlands. *Global Biogeochemical Cycles*, 11(4), 495-506.
- Reddy, K. R., DeLaune, R. D., 2008. *Biogeochemistry of wetlands: science and applications*. Crc Press.
- Roulet, N., Moore, T., Bubier, J., Lafleur, P., 1992. Northern fens: Methane flux and climate change. *Tellus*, 44B, 100-105.
- Roulet, N. T., Barrie, L., Glooschenko, W. A., Harriss, R. C., MacPherson, J. I., Moore, T. R., et al., 1992. Northern Wetlands Study (NOWES): an assessment of the role of northern wetlands in the exchange of atmospheric trace gases. *Ecological Bulletins*, 77-85.
- Sage, R. F., Kubien, D. S., 2007. The temperature response of C₃ and C₄ photosynthesis. *Plant, Cell & Environment*, 30(9), 1086-1106.
- Sage, R. F., 2002. Variation in the *k_{cat}* of Rubisco in C₃ and C₄ plants and some implications for photosynthetic performance at high and low temperature. *Journal of Experimental Botany*, 53(369), 609-620.
- Schedlbauer, J. L., Munyon, J. W., Oberbauer, S. F., Gaiser, E. E., Starr, G., 2012. Controls on ecosystem carbon dioxide exchange in short-and long-hydroperiod Florida Everglades freshwater marshes. *Wetlands*, 32(5), 801-812.

- Sebacher, D.I., Harriss, R.C., Bartlett, K.B., Sebacher, S.M., Grice, S.S., 1986. Atmospheric methane sources: Alaskan tundra bogs, an alpine fen, and a subarctic boreal marsh. *Tellus*, 38B (1), 1-10.
- Seber, G. A. F., Wild, C.J., 2003. *Nonlinear Regression*. Hoboken, NJ: Wiley-Interscience.
- Seber, G. A. F., Wild, C. J., 2005. Computational methods for nonlinear least squares. *Nonlinear regression*, 619-660.
- Sharifi, A., Kalin, L., Hantush, M. M., Isik, S., Jordan, T. E., 2013. Carbon dynamics and export from flooded wetlands: A modeling approach. *Ecological Modelling*, 263, 196-210.
- Smith, K.A., Ball, T., Conen, F., Dobbie, K.E., Massheder, J., Rey, A. 2003. Exchange of greenhouse gases between soil and atmosphere: interactions of soil physical factors and biological processes. *European Journal of Soil Science*, 54(4), 779-791.
- Smith, N. G., Dukes, J. S., 2013. Plant respiration and photosynthesis in global-scale models: incorporating acclimation to temperature and CO₂. *Global Change Biology*, 19(1), 45-63.
- St-Hilaire, F., Wu, J., Roulet, N.T., Frohling, S.E., Lafleur, P.M., Humphreys, E.R., Arora, V., 2010. McGill wetland model: evaluation of a peatland carbon simulator developed for global assessments. *Biogeosciences*, 7: 3517–3530, doi:10.5194/bg-7-3517-2010.
- Tuittila, E. S., Vasander, H., Laine, J., 2004. Sensitivity of C Sequestration in Reintroduced Sphagnum to Water-Level Variation in a Cutaway Peatland. *Restoration Ecology*, 12(4), 483-493.
- Tukey, J. W., 1977. *Exploratory data analysis*. Menlo Park, CA: AddisonWesley.
- Valiela, I., Geist, M., McClelland, J., Tomasky, G., 2000. Nitrogen loading from watersheds to estuaries: verification of the Waquoit Bay nitrogen loading model. *Biogeochemistry*, 49(3), 277-293.
- Van Huissteden, J., Van den Bos, R. M., Marticorena Alvarez, I., 2006. Modeling the effect of water-table management on CO₂ and CH₄ fluxes from peat soils. *Netherlands Journal of Geosciences*, 85, 3–18.
- Wang, Q., Wang, C. H., Zhao, B., Ma, Z. J., Luo, Y. Q., Chen, J. K., Li, B., 2006. Effects of growing conditions on the growth of and interactions between salt marsh plants: implications for invasibility of habitats. *Biological Invasions*, 8(7), 1547-1560.
- Walter, B.P., Heimann, M., 2000. A process-based, climate-sensitive model to derive methane emissions from natural wetlands: Application to five wetland sites, sensitivity to model parameters, and climate. *Global Biogeochemical Cycles*, 14, 745– 765.
- Wania, R., Ross, I., Prentice, I. C., 2009. Integrating peatlands and permafrost into a dynamic global vegetation model: 1. Evaluation and sensitivity of physical land surface processes. *Global Biogeochemical Cycles*, 23(3).

- Waquoit bay national estuarine research reserve website, 2015
<http://www.waquoitbayreserve.org/>
- Wold, S., Sjöström, M., Eridksson, L., 2001. PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2), 109-130.
- Xing, Y., Xie, P., Yang, H., Ni, L., Wang, Y., Rong, K., 2005. Methane and carbon dioxide fluxes from a shallow hypereutrophic subtropical Lake in China. *Atmospheric Environment*, 39(30), 5532-5540.
- Yurova, A., Wolf, A., Sagerfors, J., Nilsson, M., 2007. Variations in net ecosystem exchange of carbon dioxide in a boreal mire: Modeling mechanisms linked to water table position. *Journal of Geophysical Research, Biogeosciences*, (2005–2012), 112(G2).
- Yvon-Durocher, G., Allen, A. P., Bastviken, D., Conrad, R., Gudas, C., St-Pierre, A., et al., 2014. Methane fluxes show consistent temperature dependence across microbial to ecosystem scales. *Nature*, 507(7493), 488-491.
- Zhang, Y., Li, C., Trettin, C.C., Li, H., Sun, G. 2002. An integrated model of soil, hydrology, and vegetation for carbon dynamics in wetland ecosystems. *Global Biogeochemical Cycles*, 16(4), 1 - 18, doi:10.1029/2001GB001838.

Chapter 7: Data-driven mechanistic modeling of CH₄ emission from coastal wetlands:

Application of a data-analytics framework

Abstract

Methane (CH₄) is one of the major greenhouse gasses (GHGs) emitting from the wetlands and has a significant global warming potential (GWP). In this data analysis based study, the major controls and relative linkages of the environmental and biogeochemical drivers of the instantaneous CH₄ emission fluxes from coastal wetlands were determined. Four different coastal wetlands of Waquoit Bay, MA that are subjected to a moderate ecological gradient were used as the case study sites. The mechanistic attributes of the CH₄ emission acquired from the linkages were then utilized to develop a parsimonious (minimum parameter), nonlinear, data-driven model for the prediction of the CH₄ emission fluxes. The position of the water table relative to the salt marsh surface was identified as the most dominant factors controlling CH₄ flux emissions; temperature (air and soil) and soil salinity also demonstrated high to moderate linkages depending on the soil O₂ availability. The predictive empirical models of instantaneous CH₄ fluxes were developed for two different hydrological regimes separated by the water table. The models were built as a nonlinear function of the three most important drivers (water depth relative to the marsh surface, soil temperature, and soil salinity) of wetland CH₄ emission. The models exhibited very good modeling efficiency (coefficient of determination, R²=0.79-89) and accuracy (ratio of root-mean-square error to the standard deviation of observations, RSR=0.32-0.45, and mean absolute error, MAE=0.64-1.00 nmole/m²/s) in predicting the instantaneous emissions. The predictive model can be used as an ecological engineering tool to aid the policy makers and land managers to take prompt decision regarding the tidal wetland restoration and maintenance under a current change of climate and sea level rise.

7.1 Introduction

Methane (CH_4) is one of the major greenhouse gasses (GHGs) and has a significant global warming potential (GWP) compare to carbon dioxide (CO_2). According to the 2013 IPCC report, the GWP of methane is 34 and 86 for 100 and 20 years, respectively (Stocker et al., 2013; Howarth, 2014). The atmospheric concentration of CH_4 has increased over 150% from 1750 to 2011 (Hartmann et al., 2013) and CH_4 is considered as one of the contributors to ozone layer depletion. These alarming facts warrant the development of an appropriate management protocol to limit the emission rate and preserve the carbon stored beneath the soil. The wetland is regarded as the primary source of natural CH_4 to the atmosphere and plays a vital role in the soil-atmospheric exchange of the GHG fluxes (CH_4 and CO_2) (Zhu et al., 2014). Although CH_4 emission from the coastal saline water wetlands were comparatively less than the freshwater wetlands because of the high saline concentration (Poffenbarger et al., 2011), the estimation and future prediction of the CH_4 fluxes were crucial for the determination of net carbon (C) balance (net C sequestration) of the coastal wetlands. The C storage and sequestration capacity of coastal and marine wetlands are often referred to as “blue carbon” (in contrast to a forest or “green carbon”). "Blue carbon" is an emerging concept for coastal management that can potentially attract much private and public investment in coastal protection and restoration (Nellemann et al., 2009; Crooks et al., 2010, 2011; Morris et al., 2012). For the successful implementation of the “blue carbon” concept, the response of CH_4 emissions to the changes in climate (e.g., temperature, radiation), sea level rise (SLR) and inundation (e.g., soil moisture, salinity) should be investigated by determining the relative control and interactions of these climatic and hydrologic factors. Moreover, a robust and handy modeling protocol is necessary as an ecological engineering tool to estimate and predict the CH_4 emissions from the coastal wetlands for the

quantification of the ecosystem scale CH₄ budget and net C balance, which would leverage the ongoing effort for setting up GHG offset methodology for wetland restoration and maintenance.

CH₄ emissions from the coastal saline water wetlands are the function of CH₄ production in an anaerobic condition and CH₄ oxidation in aerobic condition. The entire emission mechanism is controlled by a set of biogeochemical processes mediated by ambient climate, hydrology, and soil properties. The intricate processes associated with the CH₄ emission are not independent of each other; rather a complex interaction of all the controls influence the system and set the rate of CH₄ emission from the wetlands. However, the driving factors need to be ranked based on their quantitative linkages with CH₄ emissions for the development of an optimum predictive model that will incorporate only necessary drivers to limit the number of input variables to ensure the model parsimony.

The available process-based models for the prediction of wetland greenhouse gas (GHG) fluxes were already mentioned in Chapter 6. The utility of the data-driven models compared to the available process-based models were also discussed in Chapter 6 along with a comprehensive summary of the currently available data-informed explanatory and predictive models. Moreover, the usefulness of the “big ecological data” in describing and developing wetland GHG flux models was also highlighted in the previous chapter. The systematic data analytics framework explained in chapter 4, 5 and 6 that leverage the relative linkages and pattern identification of the major drivers of terrestrial and wetland CO₂ fluxes were also applied for the determination of the relative linkages of wetland CH₄ fluxes.

The objectives of this study were to investigate and identify the dominant controls of wetland CH₄ emission fluxes and determine their relative linkages, interrelation patterns and similarity with various hydro-climatic, biogeochemical and environmental drivers by using the systematic multivariate data-analytics and informatics framework. The knowledge of similarity

and interrelation patterns of the dominant controls were then utilized to develop a data-driven, parsimonious (minimum parameter) non-linear (e.g., power-law) model for the prediction of coastal wetland CH₄ emission fluxes.

7.2 Materials and methods

7.2.1 Study sites and data set

The same study wetland ponds (Sage lot pond, Eel pond, great pond and Hamblin pond) that were used for the development of the daytime CO₂ exchange model in chapter 6 were used in this study to develop the predictive model of the CH₄ emission fluxes (Figure 6.1) under the NOAA-NERRS collaborative project (Waquoit Bay website, 2015).

Instantaneous CH₄ emission fluxes and corresponding biogeochemical, hydrological and climatic variables were subsampled from the collected dataset for different days between June to October, 2013 (4 different days in June, 2 days in July, 5 days in August, 1 day in September, and 1 day in October, and 2 days in November) (Figure 7.1) in the low marsh zone of the four study wetlands. The data collection days represent growing summer and fall seasons (June to November) of coastal salt marshes. Several instantaneous measurements were carried out in each day at different clock hours to cover the temporal variability, and both low and high tide conditions. The complete dataset included net CH₄ emission fluxes from the salt marshes and seven hydrological, biogeochemical and climatic variables (Table 7.1, Figure 7.1): CH₄ emission flux (F_{CH_4} , nmol/m²/s), air temperature (AT, °C), soil temperature (ST, °C), photosynthetic active radiation (PAR, μmol/m²/s), water depth related to marsh surface (WD, m), soil salinity (pore water) (SS, ppt), soil moisture (SM, %) and soil pH (pH). The collected climatic and environmental variables cover potential mechanistic processes/groups (temperature group: AT

and ST; radiation group: PAR; hydrology group: WD and SM; and soil biogeochemical group: SS and pH) associated with the methane flux emissions.

The standard flux sign convention, a negative sign represents emission was followed for the CH₄ fluxes. Moreover, a positive sign was used when water depth is above the salt marsh surface and –ve sign was used when water depth is below the surface. The measurement procedures of the CH₄ flux and other variables considered in this study were discussed in chapter 7 (see section 6.2.1)

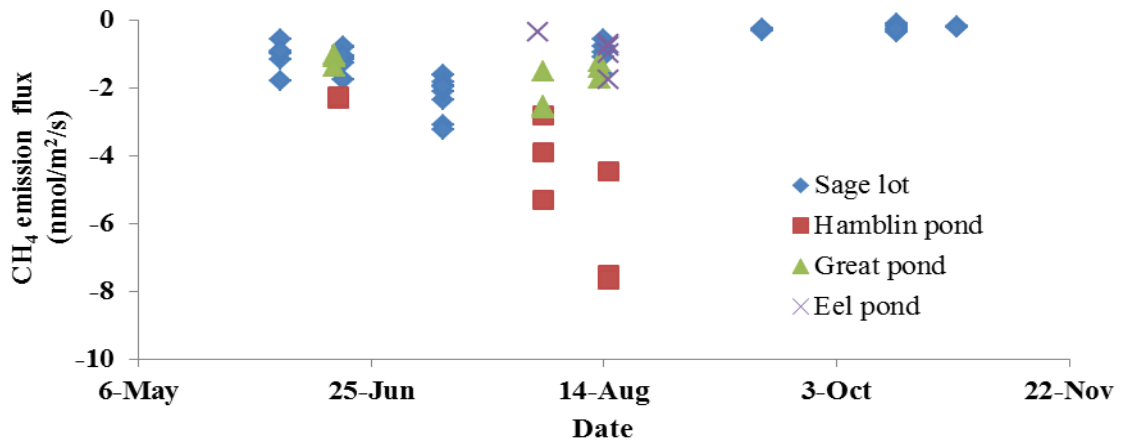


Figure 7.1: Measured CH₄ emission fluxes for the different days and diurnal hours of 2013 at the four study ponds of Waquoit Bay, MA. The negative values of the fluxes indicate emission from the marsh surface.

7.2.2 Data preparation and data statistics

The data preparation and outlier removing techniques that were applied in this study were discussed in the previous chapter (see section 6.2.2). However, in the 1st data treatment step, the linear regression coefficient, r^2 of 0.80 was used to linearly convert instantaneous CH₄ concentrations to fluxes as opposed to the r^2 value of 0.90 for F_{CO₂}. The r^2 of 0.80 was set as a threshold for the CH₄ fluxes to retain as much as possible flux data while preserving the data quality. The final dataset contains 66 observations from four study ponds (forty-four measurements from the Sage Lot, eight observations from the Hamblin pond, five from the Eel

Table 7.1: Data statistics (means, standard deviations, maximum, and minimum) of the observed (66 observations) climatic, environmental, and biological variables for the study wetlands.

Variable	PAR ($\mu\text{mol}/\text{m}^2/\text{s}$)	WD (m)	AT ($^{\circ}\text{C}$)	ST ($^{\circ}\text{C}$)	pH	SS (ppt)	SM (%)	F_{CO_4} ($\text{nmol}/\text{m}^2/\text{s}$)
Data	1389.4 \pm 639.5	0.09 \pm 0.1	28.1 \pm 6.3	20.5 \pm 4.1	6.9 \pm 0.32	29.2 \pm 3.8	63.3 \pm 3.5	1.53 \pm 1.5
Statistics	Max=2080.4 Min=115.6	Max=0.33 Min=-0.07	Max=37.6 Min=12.8	Max=27.8 Min=8.9	Max=7.9 Min=6.1	Max=34 Min=20	Max=68.7 Min=51.2	Max=7.7 Min=0.096

Note: PAR, WD, AT, ST, pH, SS, SM, and F_{CH_4} refer, respectively, to photosynthetic active radiation, water depth relative to water marsh surface, air temperature, soil temperature, soil pH, soil salinity, soil moisture, and CH_4 emissions from the salt marshes. Positive sign of WD means water table above the salt marshes. The unit of each variable was given in the parenthesis.

pond, and nine observations from the Great pond). Figure 7.1 showed the range of data collection days where each day had multiple data points at different daytime hours.

The observed methane emissions from the wetlands were not high; the growing season average of the collected F_{CH_4} for the study wetlands was 1.53 nmol/m²/s (Table 7.1). The maximum emission rate was 7.67 nmol/m²/s as measured in the Hamblin pond (Figure 7.1). Overall, the Hamblin pond demonstrated higher CH₄ emissions compare to the other study ponds in Waquoit bay system. The spread of the CH₄ fluxes as seen in Figure 7.1 suggested the incorporation of different emission regimes (low and high emission) of CH₄ fluxes across the selected study wetlands. The ambient air temperature was as high as 37.6 °C with a corresponding high soil temperature. The mean pH was found to be neutral, and that suggest high productivity of the salt marshes and adjacent plant communities. Soil salinity (pore water) was considerably high over the growing period (SS ranged from 20-34). The measured PAR covered both cloudless (when PAR is high), and cloudy and evening environments (Low PAR) of the study wetlands. The positive value of mean WD indicated the presence of a dominant anoxic state in the study sites.

7.2.3 Data analysis and predictive modeling

The biogeochemical activity and CH₄ emission from the coastal salt marshes largely depend on the oxygen availability in the sediment pores. Soil water level determines this aerobic-anaerobic state of the salt marshes. Since the coastal wetlands are subjected to diurnal flood and ebb tides, the potential linkages between the different climatic variables and F_{CH_4} due to the fluctuating water table were investigated for two conditions.

Condition 1: Water above the salt marsh surface (WD>0)

Condition 2: Water below the salt marsh surface (WD<0).

The final dataset was divided based on the above-mentioned conditions, and all the multivariate statistical analysis methodology described below was performed discretely for these two settings. Out of the 66 observation panels, 51 observations that comprise all the four study sites had water above the salt marsh surface. The rest of the 16 observation panels followed the condition 2 subset ($WD < 0$).

Data-analytics framework and modeling

The data analytics framework (Figure 6.3) discussed in the chapter 6 was applied to determine the relative controls and linkages of the climatic and environmental variables with wetland F_{CH_4} (please see section 6.2.3). The detail description of the framework is given in chapter 4 (see section 4.2.3). The most important and dominant drivers controlling the CH_4 emission fluxes in the coastal salt marshes were identified based on the data-analytics framework and mechanistic judgment. These selected variables were then utilized to develop a power-law based non-linear parsimonious (i.e., minimum parameter set) model (Equation 1) of F_{CH_4} for the coastal salt marshes. The details of the power-law based cross-validated modeling were discussed in chapter 6 (please see section 6.2.3)

$$F_{CH_4} = 10^a \prod_{i=1} X_i^{b_i} \quad (1)$$

where a and b = model parameters (a is the scaling factor, and b represented shape factors), X_i = independent variables (important process drivers of F_{CH_4}).

7.3 Results

7.3.1 Correlation analysis

Condition 1: Water above the salt marsh surface (WD>0)

Since data was log-transformed and standardized prior to the analysis, only absolute values of the correlations were considered for result interpretation. The CH₄ emission flux, F_{CH₄} showed a high correlation with temperature (AT, ST) variables ($|r|=0.85-0.88$) while PAR, WD, and SS showed a moderate correlation ($|r|=0.41-0.65$) (Table 7.2). However, soil moisture (SM), a surrogate indicator of soil O₂ availability, and pH were poorly correlated with the F_{CH₄} ($|r|=0.03-0.04$). All the correlations were significant at 95% level of confidence. The radiation and temperature-related variables demonstrated significant inter-variable correlations (AT-ST, PAR-AT, and PAR-ST) as obtained from the triangular correlation matrix (not shown) implying the presence of a multicollinearity in the independent variable data set.

Condition 2: Water below the salt marsh surface (WD<0)

The hydrologic variable, WD showed a high correlation with the F_{CH₄} for the low tide condition ($|r|=0.77$). The other hydrology related variable, SM was weakly correlated ($|r|=0.22$) with F_{CH₄}. The temperature and radiation variables did not demonstrate notable correlations with methane ($|r|=0.10-0.20$). Similarly, soil redox variables (SS and pH) showed low correlations. The presence of the multi-collinear structure in the independent variable data set was apparent since moderate correlations were observed between AT-ST, SS-WD, and PAR-pH (not shown).

For both of the hydrologic conditions (WD>0 and WD<0), the dominant influence of water depth on CH₄ emission was observed; however, temperature and radiation variables demonstrated two different correlation strengths for the varying hydrologic conditions.

Table 7.2: Pearson correlation coefficients between the CH₄ emission fluxes (F_{CH₄}) and the corresponding climatic and environmental variables for the study area for both WD>0 (water table above the soil marsh) and WD<0 (water table below the soil marsh) conditions.

	PAR	WD	AT	ST	pH	SS	SM
Condition 1 (WD>0)	0.65	-0.43	0.88	0.85	-0.03	-0.41	0.04
Condition 2 (WD<0)	0.10	0.77	-0.19	0.20	-0.15	0.06	-0.22

Note: PAR, WD, AT, ST, pH, SS, SM, and F_{CH₄} refer, respectively, to photosynthetic active radiation, water depth relative to water marsh surface, air temperature, soil temperature, soil pH, soil salinity, soil moisture, and CH₄ emissions from the salt marshes.

Condition 1(WD>0)

1st two principal components (PC) explained 63.6 % of the total data variances for all the study ponds. The loading of each variable on the 1st two orthogonal components were presented through biplot (Figure 7.2a). The relative orientation and length of the participatory variables in the biplot revealed the existence of two different variable clusters for the anoxic condition. The non-orthogonal (0 or 180 degree) orientations between the F_{CH₄} and the group of PAR-AT-ST suggested their higher interrelationship and linkages (all of these variables loaded highly on 1st PC). The relative position of WD and F_{CH₄} in the 2-D plane indicated their moderate linkages and control. A nearly orthogonal (90 degree) orientation was observed between F_{CH₄} and group of pH-SM; suggesting their weak linkages (pH loaded highly on 2nd PC while F_{CH₄} loaded high on 1st PC and SM loaded very weakly on both of the PCs). Soil salinity (SS) loaded moderately to high on the both PCs; the relative orientation of SS with F_{CH₄} suggested a moderate CH₄ flux linkages. Orthogonal loading pattern between the temperature and radiation variables (PAR, AT, and ST) and the hydrology-redox variables (WD, pH, SS, and SM) indicated the presence of two process clusters that was previously hidden in the data. The 1st cluster (radiation-temperature) was driven by climate and highly linked with the F_{CH₄}. The 2nd cluster (WD, SM, pH and SS) resembled wetland hydrology and biogeochemistry, which were weak to moderately correlated with F_{CH₄}.

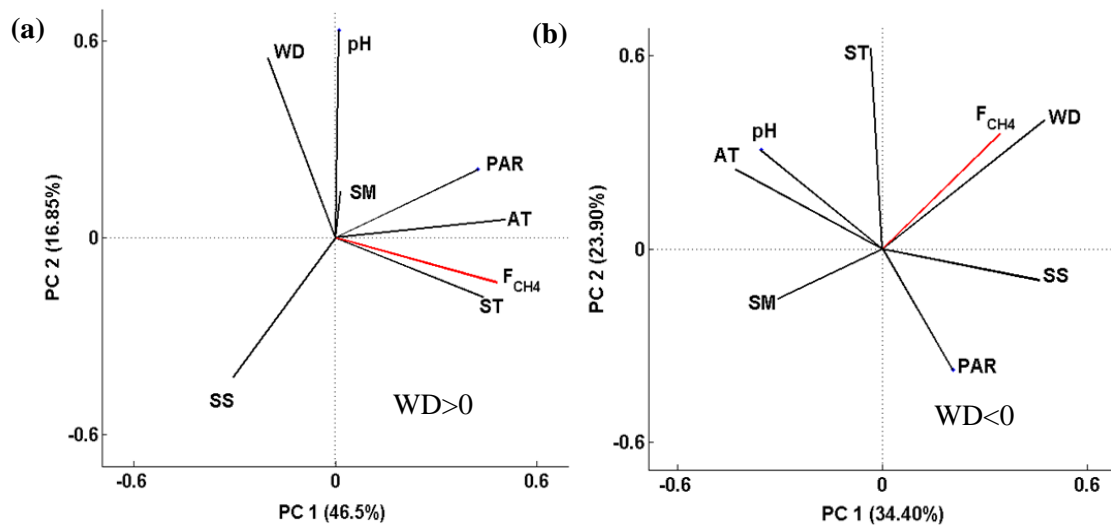


Figure 7.2: Biplots obtained from the loadings of the 1st two principal components (PC) showing the groupings and interrelation patterns of the climatic, environmental, and biological variables for both (a) WD>0 and (b) WD<0 conditions. Percent variance explained by each PC is shown in parenthesis. PAR, WD, AT, ST, pH, SS, SM, and F_{CH4} refer, respectively, to photosynthetic active radiation, water depth relative to water marsh surface, air temperature, soil temperature, soil pH, soil salinity, soil moisture, and CH₄ from the salt marshes.

Condition 2 (WD<0)

The 1st two PC's explained 58.3% of the total data variances (Figure 7.2b) of the analysis data-matrix. Contrary to the observations of condition 1, no apparent cluster of variables emerged from the biplot for the condition 2. Hydrology variables (WD and SM) were highly linked with F_{CH4} because of their non-orthogonal orientation. However, WD had the most CH₄ linkage than SM; WD loaded heavily on both PCs similar to the loading of F_{CH4} while SM loaded moderately on both PCs. The orientation of SS and ST relative to F_{CH4} suggested a moderate CH₄ flux linkages. Since AT and pH were orthogonally linked with F_{CH4}, they had relatively lower linkages.

7.3.3 Extracted orthogonal latent factors

Condition 1(WD>0)

Three latent factors optimally explained the overall data variances for the water level above the salt marsh condition (Table 7.3); Eigen value>1 based optimization while extracting the factors ensured the systematic consideration of most of the data variances (78% of the data variances

were explained by the 1st three factors) for process description. The factors were ranked on the basis of the percent data explained by each factor; 1st factor explained 46.38%, 2nd factor explained 17.30%, and 3rd factor explained 14.29 % of the total data variances. Since, log-transformed Z-scores were used for the analysis, only the magnitudes of the loadings were considered for the explanation of the FA results.

Table 7.3: Extracted latent factors from the factor analysis for the study wetlands for both WD>0 and WD<0 (shaded area in the table) conditions.

Note: PAR, WD, AT, ST, pH, SS, SM, and F_{CO2} refer, respectively, to photosynthetic active

Condition	Factors	PAR	WD	AT	ST	pH	SS	SM	F _{CO2}
WD>0	1	0.48	-0.38	0.83	0.94	-0.10	-0.23	-0.03	0.87
	2	0.84	0.09	0.41	0.13	0.05	-0.24	0.28	0.23
	3	-0.23	0.01	-0.36	0.02	-0.27	0.83	0.22	-0.17
WD<0	1	0.11	0.76	-0.21	0.17	-0.15	0.06	-0.21	0.99
	2	0.07	0.54	-0.75	-0.05	-0.29	0.83	-0.30	0.01
	3	-0.38	0.30	0.48	0.98	0.23	0.04	-0.11	0.03

radiation, water depth relative to water marsh surface, air temperature, soil temperature, soil pH, soil salinity, soil moisture, and CH₄ emissions from the salt marshes. Bold values indicate most important variables in each factor.

Temperature variables (AT, ST) loaded highly (0.83-0.94) on 1st factor along with a very high loading (correlation between a factor and a variable) of F_{CH4} (0.87) on this factor; suggesting a strong methane flux linkages. Moderate loading of WD in factor 1 (-0.38) showed the existence of moderate hydrological control on F_{CH4}. PAR loaded moderately on the 1st factor (0.48) and very highly on the 2nd factor (0.84); however, loading of F_{CH4} on the 2nd factor was not high (0.23) suggesting a moderate linkage with PAR. SS loaded highly on the 3rd factor (0.83) with a low loading of F_{CH4} (-0.17) indicating a weak to moderate dominance of SS on F_{CH4}. The other participatory variables (pH and SM) did not load significantly on the extracted factors. The lower loading of these variables indicated weak linkages and correlation with the F_{CH4} in the flooding condition.

Condition 2 (WD<0)

Similar to the condition 1, three latent factors were extracted for the optimal interpretation of the overall data variances. Three extracted factors explained 71.2% of the variance where the 1st factor explained 34.4%, the 2nd factor explained 23.90%, and the 3rd factor explained 12.90% of the total variations.

Water depth (WD) loaded highly (0.76) on the 1st factor with a very high loading of F_{CH_4} on this factor (0.99); indicating a strong relationship between WD and F_{CH_4} at WD<0 condition (Table 7.3). Although loading of AT and SS on factor 2 were high (-0.75 and 0.83), very small loading of F_{CH_4} (0.01) on factor 2 suggested their weak to moderate influences. Similarly, ST loaded profoundly (0.98) on factor 3 along with a low loading of F_{CH_4} (0.03); inferring weak linkage. PAR, SM, and pH did not exhibit any notable loading on any on the extracted factors in this condition.

7.3.4 Partial least squares modeling in non-linear domain

The correlation analysis, PCA, and FA facilitated to identify the dominant variables that mostly triggered CH_4 emissions from the salt marsh dominated wetlands at varying hydrological conditions. However, these series of statistical investigations did not shed light on the relative magnitude of the CH_4 flux linkages and percent variance explained by the different process groups. To quantify the relative linkages for individual variables as well as for different process groups, nonlinear partial least squares models were developed for both aerobic and anaerobic conditions with the log-transformed, normalized data of the dependent (F_{CH_4}) and independent (PAR, AT, ST, SM, WD, SS, and pH) variables. The AIC optimization (minimum AIC) and R^2 (maximum R^2) led to the selection of 1st three PLS components for both WD>0 and WD<0 conditions (Figure 7.3) while developing the PLSR models. The individual quantitative linkages were calculated by dividing the highest weighted coefficient (BETA) in the regression by the

other estimated coefficients. The optimized PLSR coefficients (BETA) were also leveraged to determine the percent of modeled F_{CH_4} explained by the different process groups (i.e., temperature, radiation, hydrology, and biogeochemical). Since regression weights BETA contain positive and negative signs, to calculate the component/group wise quantitative linkages, the combined weights of the “radiation” group (β_R), “temperature” group (β_T), “hydrology” group (β_H) and “biogeochemical” group (β_B) were computed, respectively based on following equations:

$$\beta_R = \sqrt{\beta_{PAR}^2}; \beta_T = \sqrt{\beta_{AT}^2 + \beta_{ST}^2}; \beta_H = \sqrt{\beta_{WD}^2 + \beta_{SM}^2}; \text{ and } \beta_B = \sqrt{\beta_{SS}^2 + \beta_{pH}^2}$$

The percent influence of these groups on F_{CH_4} was then obtained by calculating percent explained

by them, as: $\frac{\beta_R}{\beta_R + \beta_T + \beta_H + \beta_B} * 100$ (Influence of radiation factor)

$$\frac{\beta_T}{\beta_R + \beta_T + \beta_H + \beta_B} * 100 \text{ (Influence of temperature factor); } \frac{\beta_H}{\beta_R + \beta_T + \beta_H + \beta_B} * 100$$

(Influence of hydrology factor); and $\frac{\beta_B}{\beta_R + \beta_T + \beta_H + \beta_B} * 100$ (Influence of biogeochemical factor).

Condition 1(WD>0)

The developed 3-component PLSR model was statistically significant and based on the modeling efficiency and accuracy, the model could be labeled as “good” ($F_{2,48}=130.45$, $p\text{-value}<0.0001$, $R^2=0.84$, $RSR=0.39$, and $MAE=0.32$) (Table 7.4). The observed high modeling efficiency indicated a presence of dominant non-linear liaison between the dependent and independent variables. The estimated model coefficients, BETA were reported in Table 7.4. Temperature variable (AT and ST) and hydrology variable WD received most of the weights (ST got the

highest weight) in the regression, which is consistent with the observation of PCA and FA. Overall, ST exhibited 1.3- 2.2 times stronger linkages with F_{CH_4} than that of WD and AT; 7.2-7.7 times stronger linkages than that of PAR and SM; 11.1-13.8 times stronger linkages than that of SS and pH. Regarding the group wise influence of the variables on F_{CH_4} , temperature and hydrology group explained much of the modeled F_{CH_4} (63.5% and 23.8%) while the radiation (6.9%) and biogeochemical (5.6%) group together explained only 12% of the variances.

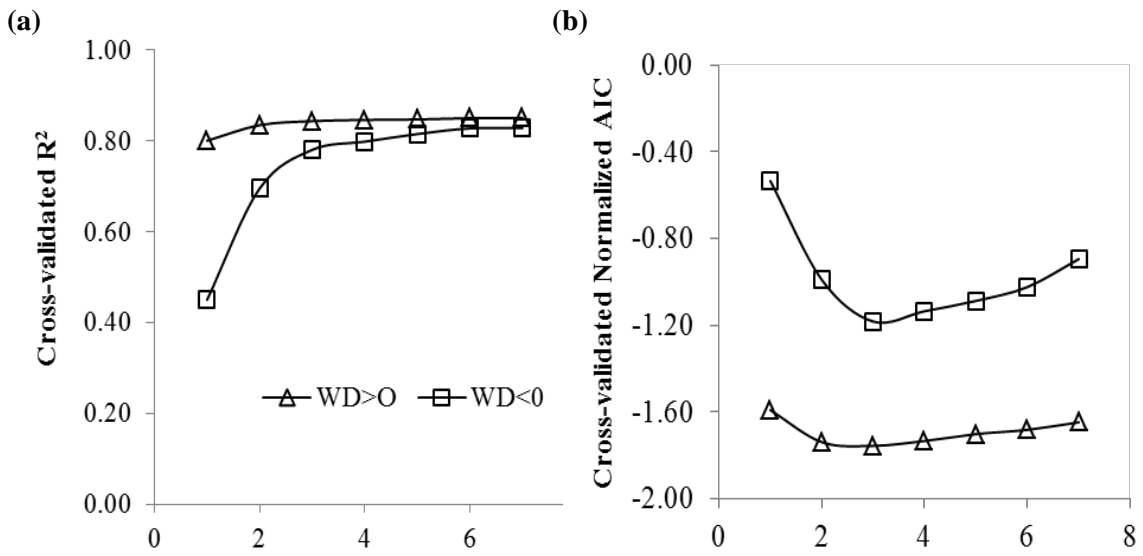


Figure 7.3: Plot of 10-fold cross-validated (a) fitting efficiency, R^2 and (b) Normalized AIC for both WD>0 and WD<0 conditions of instantaneous F_{CH_4} with the number of incorporated partial least squares (PLS) components.

Condition 2(WD<0)

Similar to the condition 1, PLSR model developed for condition 2 efficiently described the data variances and the model was statistically significant ($F_{2,13}=23.01$, p -value<0.0001, $R^2=0.78$, $RSR=0.45$, and $MAE=0.38$) (Table 7.4). Based on our labeling criteria, the model could be ranked as “good”. Water depth (WD) was the most influential variables in the regression followed by a moderate dominance of SS and PAR. Coefficients of the temperature variables (AT and ST) were low that suggest weak linkages. Overall, WD showed 1.8 times stronger linkages

with F_{CH_4} than that of SS; 5 times stronger linkages than that of PAR; 9.2 times stronger linkages than that of pH; and 18.9-33.5 times linkages than that of AT, ST, and SM. The hydrology group (54.55%) and the biogeochemical group (30.92%) collectively explained almost 86 % of the variances of model F_{CH_4} when water is below the salt marsh surface while rest of the variances were explained by the temperature and radiation components.

Table 7.4: Coefficients (BETA) of the log-transformed, normalized (dimensionless) 3-component (optimized) PLSR models of the F_{CH_4} for both $WD>0$ and $WD<0$ conditions.

Variable	PAR	WD	AT	ST	pH	SS	SM	R ²	RSR	MAE
Condition 1 ($WD>0$)	0.06	-0.21	0.35	0.45	0.03	-0.04	0.06	0.84	0.39	0.32
Condition 2 ($WD<0$)	0.20	1.02	-0.05	-0.04	-0.11	-0.57	0.03	0.78	0.45	0.38

Note: PAR, WD, AT, ST, pH, SS, SM, and F_{CO_2} refer, respectively, to photosynthetic active radiation, water depth relative to water marsh surface, air temperature, soil temperature, soil pH, soil salinity, soil moisture, and CH_4 emissions from the salt marshes. R^2 , RSR, and MAE refer, respectively, to coefficient of determination, ratio of root-mean-square error to the standard deviation of observations, and mean absolute error.

7.3.5 Predictive model development

Choice of variables for predictive modeling

The 4-layer multivariate statistical analysis provided a clear insight and direction in selecting the most dominant and mechanistically viable drivers for the development of the parsimonious predictive model. Although different response-feedback loops were observed under different hydrologic conditions, a common pattern related to the control of variables on F_{CH_4} emerged. Water depth played the most important role in CH_4 emission in concert with the temperature and hydrology mediated biogeochemical activity. To incorporate these mechanistic attributes in the predictive model, the predictor variable set should include at least one variable from each of the process group (temperature, hydrology, and biogeochemistry, radiation,) given that the selected variables have direct mechanistic link with the F_{CH_4} . As such, based on the relative controls of the drivers of each group ST, (represents ambient temperature and climate), WD (hydrology group), and SS (biogeochemical group) were considered as the predictor variables for the parsimonious

nonlinear model development for both $WD > 0$ and $WD < 0$ conditions. Although radiation related variable, PAR showed weak to moderate linkages with F_{CH_4} , it was not included in the predictor variable set because CH_4 emission was not directly driven by PAR. Since AT was used to calculate the F_{CH_4} exchanges and ST showed higher linkages than AT, ST was considered in the predictor set as a temperature input. Finally, the predictor variable set was consisted of ST, WD, and SS; leading to the development of four parameter model (one scaling factor, a and three shape factors for WD, ST, and SS, respectively) (Equation 1 for $i=1$).

Since WD had both positive and negative value depending on the location of the water table, only absolute values of WD was considered while developing the predictive model for $WD < 0$ condition. Finally two models were developed for $WD > 0$ and $WD < 0$ to predict the CH_4 emission from the salt marshes by using a Monte-Carlo simulation (10000 iterations for training and testing) based power law (i.e., nonlinear) function (Equation 1).

Predictive model simulation

WD > 0 condition

The model showed “good” agreement with the observed data for training (i.e., calibration), and testing (i.e., validation) (mean calibration $R^2=0.79$; mean validation $R^2= 0.77$). The histograms of the estimated parameters plotted with the results of 10000 simulations indicated that the parameters were normally distributed (Figure 7.4). The standard deviation of the each parameter from the mean was quantified (Figure 7.4); the standard deviations indicated the uncertainty of the parameters for the model simulation. The standard deviation varied from 0.02-0.35 for all the parameters. The mean p-value of the parameters ensured the statistical significance of the parameters in the nonlinear regression (mean p-value <0.005) (Table 7.5). Finally, instantaneous F_{CH_4} for $WD > 0$ condition were predicted as a function of instantaneous ST, WD, and SS from the mean (mean over 10000 iterations) of the estimated parameters (Table 7.5) and final model

showed overall high statistical significance, and modelling efficiency and accuracy ($F_{3,47}=63.84$, p -value <0.0001 , $R^2=0.79$, $RSR=0.45$, $MAE=1.00$ nmole/m²/s). Figure 7.7 (a) showed the 1:1 plot of the predictions with the observed data for all the study ponds.

WD<0 condition

The model for the no-flood condition demonstrated “very good”, and “good” modeling efficiency during training, and testing respectively (Calibration $R^2=0.89$, validation $R^2=0.71$). Although estimated parameters were relatively normally distributed (Figure 7.5), none of the parameters but WD were statistically significant at the 95% level of significance (Table 7.5); that implied model developed from only WD could be sufficient to predict the CH₄ fluxes with a comparatively similar accuracy. A very high dominance and control of WD on F_{CH_4} as unraveled from the data-analytics framework nullified the influence of the other predictor variables in the model. From the mean of the estimated parameters, CH₄ fluxes were predicted and model showed “very good” agreement with the observations and the entire model was statistically significant ($F_{3,12}=36.43$, P -value <0.0001 , $R^2=0.89$, $RSR=0.32$, $MAE=0.64$ nmole/m²/s) (Table 7.5). The 1 to 1 plot of the observed and predicted F_{CH_4} showed a very good agreement except for one very low value of the observation (Figure 7.7b). Since the parameter corresponds to WD was the only statistically significant parameter in the regression, another nonlinear predictive model of CH₄ fluxes was developed for the WD<0 condition by considering only WD as the independent variable. The cross-validated model (WD only) calibration and validation showed “very good” modeling efficiency (Calibration $R^2=0.86$, validation $R^2=0.70$). The model parameters were normally distributed (Figure 7.6) and the standard deviation of the scaling factor a and the shape factor associated with WD were 0.21 and 0.16, respectively. The final model based on the mean of the estimated parameters also demonstrated “very good” performance in predicting instantaneous CH₄ fluxes ($F_{3,12}=27.73$, p -value <0.0001 , $R^2=0.86$, $RSR=0.37$, $MAE=0.66$ nmole/m²/s) (Table

7.5) and the parameters were statistically significant (p-value<0.05). The 1 to 1 plot (Figure 7.7c) was nearly as good as Figure 7.7b.

The choice about the selection of the model equation for the WD<0 condition largely depends on the data availability and users discretion. Although, all the parameters of the 1st model (WD<0) were not statistically significant (the overall model is statistically significant based on F-statistics), it is recommended to use the 1st model (equation 2) because it incorporated most fundamental mechanistic attributes of CH₄ emissions for the Waquoit Bay. However, the instantaneous CH₄ emission fluxes of coastal salt marshes could be predicted from any of the following equations (Equation 2 and 3)

$$F_{CH_4} = \begin{cases} 10^{-3.12} ST^{3.54} WD^{-0.19} SS^{-1.17} & \text{when water is above the marsh surface} \\ 10^{1.55} ST^{-0.95} WD^{1.10} SS^{1.20} & \text{when water is below the marsh surface} \end{cases} \quad (2)$$

$$F_{CH_4} = \begin{cases} 10^{-3.12} ST^{3.54} WD^{-0.19} SS^{-1.17} & \text{when water is above the marsh surface} \\ 10^{2.24} WD^{1.25} & \text{when water is below the marsh surface} \end{cases} \quad (3)$$

Table 7.5: The estimated parameters of the predicted model (obtained from the mean of the parameters through 10000 iterations) of F_{CH₄} for the Waquoit bay, MA wetlands. The mean p-values of the estimated parameters are given in parenthesis

Condition	Scaling factor, a	Shape factor, b			R ²	RSR	MAE (nmol/m ² /s)
		ST	WD	SS			
WD>0	-3.12 (<0.05)	3.54 (<0.05)	-0.19 (<0.05)	-1.17 (<0.05)	0.79	0.45	1.00
WD<0 (3 variable model)	1.55 (>0.05)	-0.95 (>0.05)	1.10 (<0.05)	1.20 (>0.05)	0.89	0.32	0.64
WD<0 (1 variable model)	2.24 (<0.05)		1.25 (<0.05)		0.86	0.37	0.66

Note: ST, WD, and SS refer, respectively, to soil temperature, water depth relative to the marsh surface, and soil salinity. R², RSR, and MAE refer, respectively, to coefficient of determination, ratio of root-mean-square error to the standard deviation of observations, and mean absolute error.

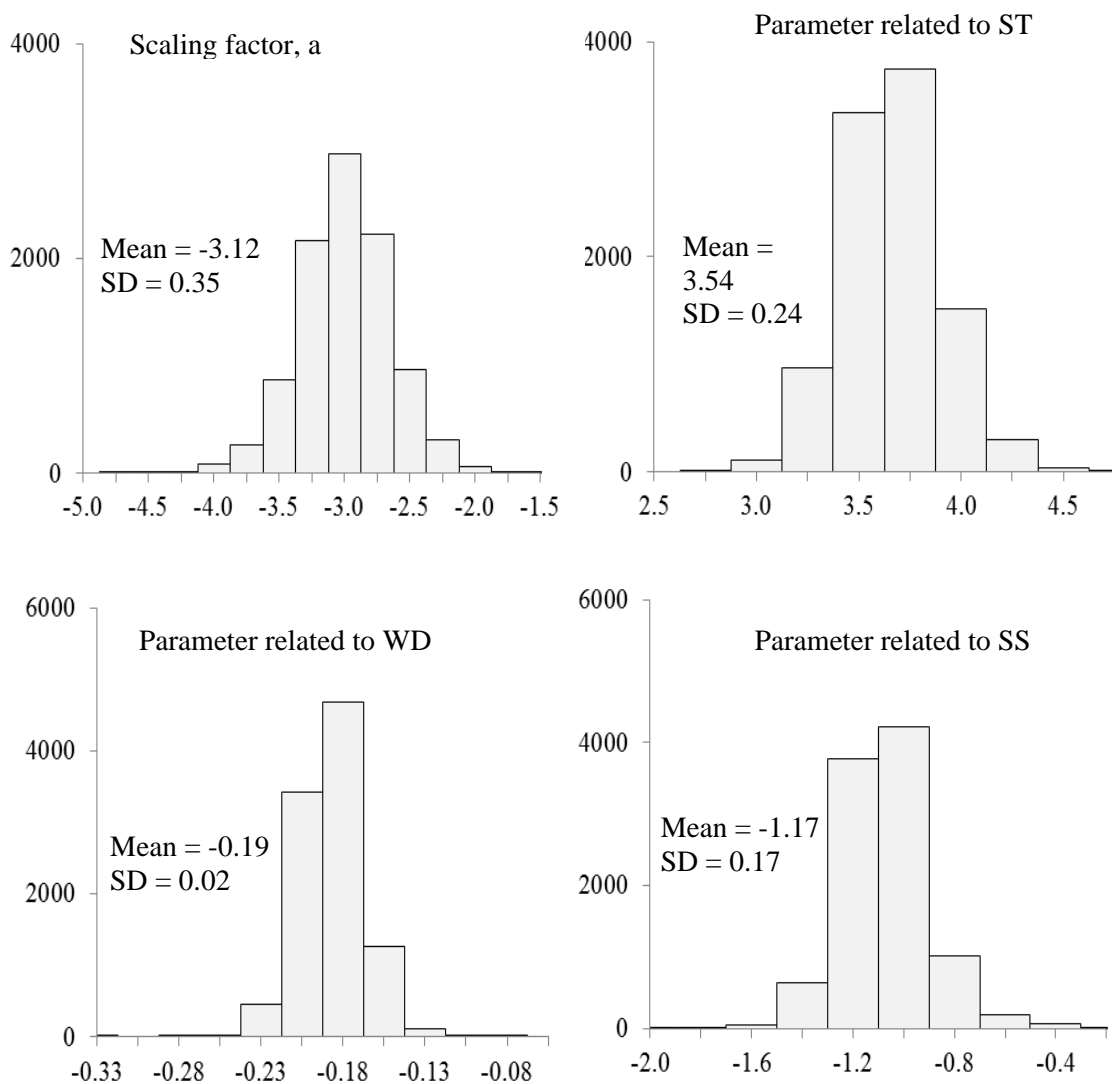


Figure 7.4: Plot shows the histogram of the estimated parameters of the predictive model for $WD > 0$ condition obtained from the Monte Carlo cross validation through 10000 iterations.

7.4 Discussion

7.4.1 Relative CH_4 flux linkages of the participatory physical drivers

The temperature and hydrology variables showed dominant control on CH_4 emission for the completely submerged condition (Figure 7.2, Table 7.3, 7.4). Soil temperature was identified as the most important driver of CH_4 emission in this condition. However, the dominance of

temperature subsided in $WD < 0$ condition as hydrology and biogeochemical group together explained more than 80% of the variances of modeled CH_4 . Position of the water table relative to the marsh surface emerged as the most dominating factor affecting F_{CH_4} .

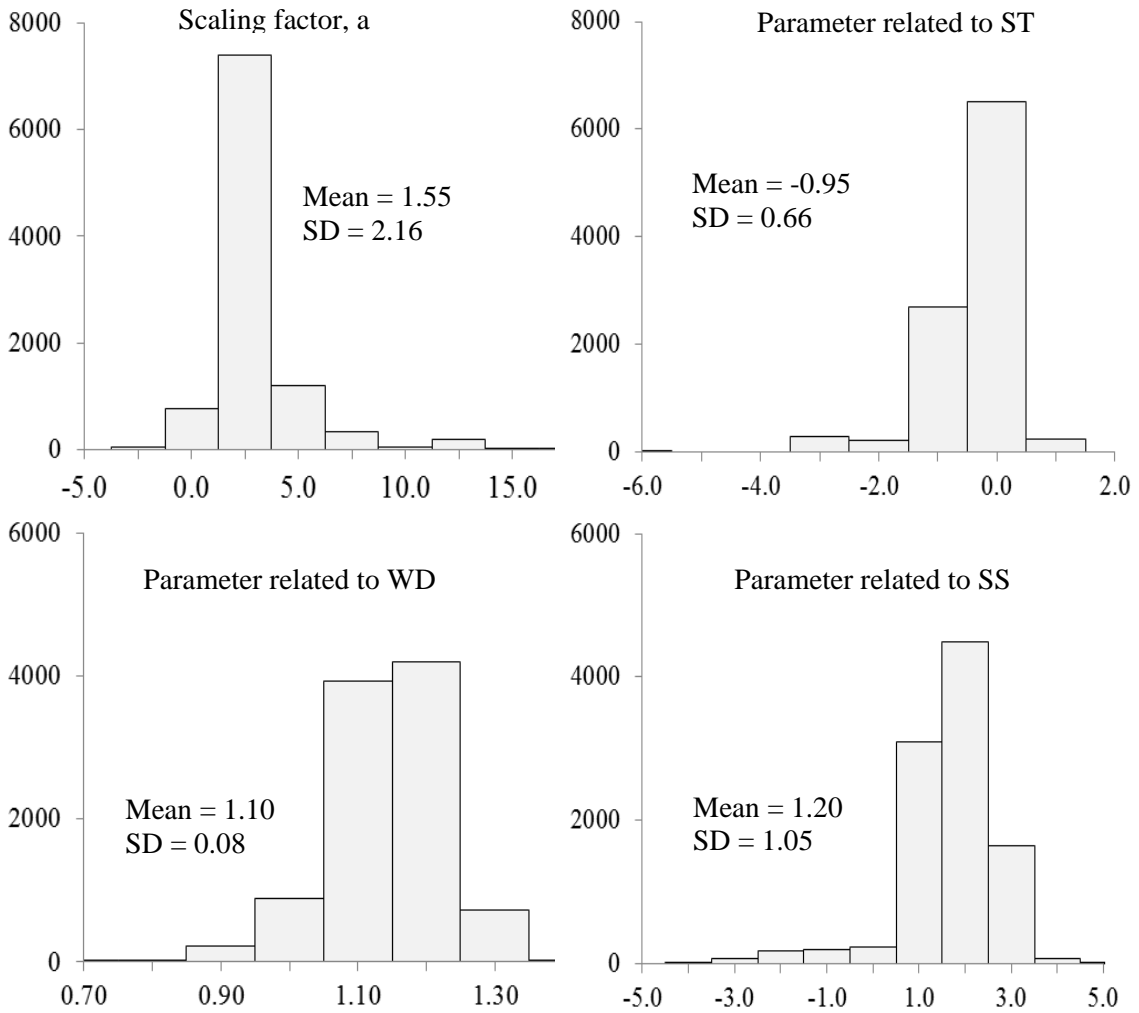


Figure 7.5: Plot shows the histogram of the estimated parameters of the 4- parameter predictive model for $WD < 0$ condition obtained from the Monte Carlo cross validation through 10000 iterations.

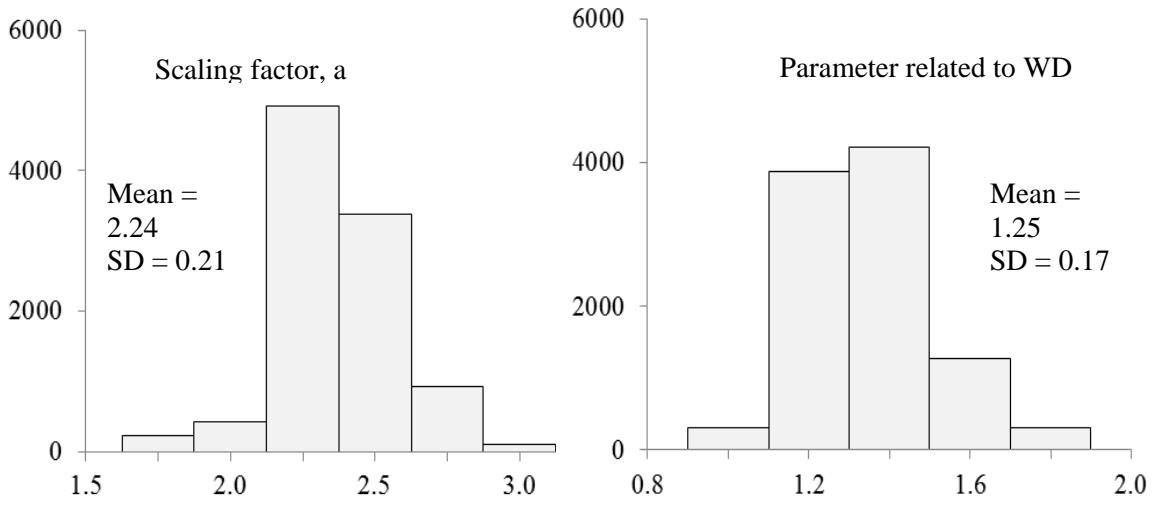


Figure 7.6: Plot shows the histogram of the estimated parameters of the 2-parameter predictive model obtained from the Monte Carlo cross validation through 10000 iterations.

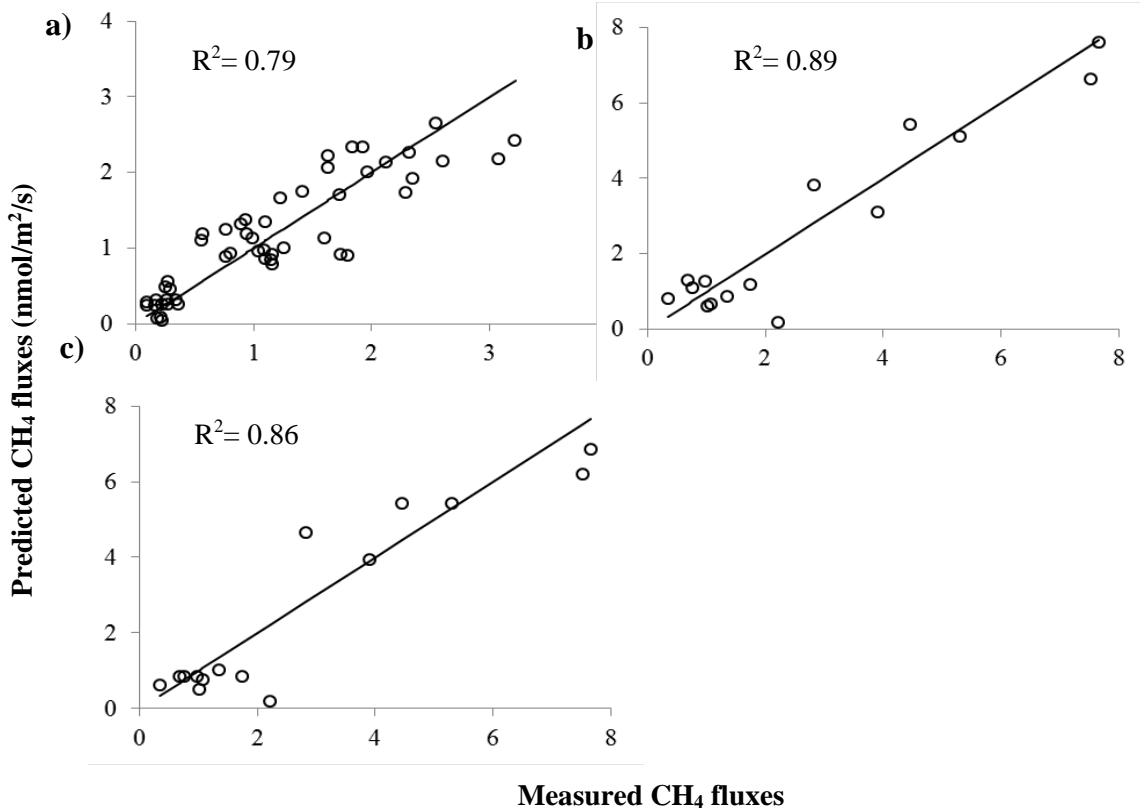


Figure 7.7: Plot of the observed versus predicted CH₄ fluxes of the coastal wetland for (a) WD > 0 condition, (b) WD < 0 condition with three variable model, and (C) WD < 0 condition for the one variable (WD only) model.

The overall higher control of water depth and soil saturation on CH₄ emission fluxes was comprehensively documented in the literature (Cui et al., 2005; Cheng et al., 2007; Bridgham et al., 2013; Moore et al., 2011; Bubier and Moore 1994). The CH₄ emission from a wetland is the function of CH₄ production (methanogenesis), CH₄ oxidation, and CH₄ emission (methanotrophy); all of these microbial-mediated processes are regulated by the water table dynamics as water table and soil saturation determine the reduced environment below the soil surface. An extremely reduced environment (negative redox potential), which is favorable for the CH₄ production and emission, occurs during the flood conditions. On the contrary, as the water subsides during ebb tide, CH₄ oxidization occurs in the anoxic zone of the soil surface (typically in the upper soil layer) that reduces the rate of CH₄ emission.

The microbial activity associated with CH₄ production is also a function of the temperature and heat (mainly soil temperature). Since water table determines the heat diffusion in the organic soil, the influence of water table position and soil temperature is closely linked (Roulet et al., 1992). The average control of temperature on the CH₄ emission for a range of ecosystems is explained through the Arrhenius equation where CH₄ emission responds exponentially with the temperature (Yvon-Durocher et al., 2014). Temperature also controls the seasonal variation of CH₄ emission. Since methane oxidation in the aerobic condition (WD<0) reduces the emission amount, the control of temperature on the net CH₄ emission diminishes as most of the methane produced in the anoxic condition in an ambient soil climate are converted to CO₂ (oxidizing). That explains the relatively lower control of temperature during the low flood condition than the flood condition scenario.

A relatively low CH₄ flux linkages of soil salinity and pH were observed for the Waquoit Bay system for WD>0 condition, and moderate influence of these variables for WD<0 condition. The optimum pH favorable for the CH₄ production is 7, which is the optimum pH level for

methanogenic bacteria (Segers, 1998). The optimum pH value is nearly equal to the average value of pH as measured in this study (Table 7.1). However, Dunfield et al. (1993) reported an optimum pH value of 6 for both the production and consumption. Moreover, there was an evidence of lack of correlation between pH and methane fluxes as reported by Bubier and Moore (1994).

Regarding the soil salinity, being a polyhaline tidal marsh the Waquoit Bay is subjected to the low CH₄ emission compare to the relatively less saline wetlands. In the sulfate-rich salt marshes (available in the salt-rich ocean water), methanogens have to compete with the sulfate-reducing bacteria for the production of CH₄ and that limit the CH₄ emission for the salinity-driven wetlands to some extent (Poffenbarger et al., 2011); however, this relationship may vary from site to site depending on the adaptation mechanism of the system (Megonigal et al., 2004). The study indicated a moderate control of SS on the CH₄ emissions for WD<0 condition. Data analytics showed that SS was not significantly linked or correlated with any other independent variables (Figure 7.2); this suggests the role of salinity on CH₄ emission could not be strongly combined with the influence of other variables. However, as SS decreases with the decrease of the water table, the observed moderate linkage of SS could due be to the WD mediated production-oxidization ration of CH₄. It could be hypothesized that due to a relatively lower influence of temperature factor in the WD<0 condition, the dominance of SS on F_{CH₄} was developed as SS varied with the change of WD.

PCA and FA indicated moderate to high methane flux linkage of PAR for WD>0 condition and weak to moderate linkage for WD<0 condition (Figure 7.2, Table 7.3). However, explanatory PLSR coefficients indicated low, and moderate control of PAR for WD>0 and WD<0, respectively (Table 7.4). Such discrepancy between the PCA and PLSR may be due to the existence of the co-linearity between PAR-ST, and PAR-AT at WD>0 condition. Since PAR, ST and AT were correlated, ST and AT shared the influence of PAR in the PLSR modeling for WD>0 condition. The moderate BETA value of PAR at WD<0 conditions may reflected the true

weight since PAR was not highly correlated with other variables in the case. Nevertheless, observed methane flux linkage of PAR could be explained from the biomass perspective. Since both aboveground and belowground productions of the plant largely depend on the photosynthesis, PAR can be considered as a surrogate of biomass. Previous studies found evidence of possible linkages between plant properties and CH₄ fluxes (Garnet et al., 2005; King and Reeburgh, 2002); however, the mechanisms driving the process is not yet understood well. Trace of “new carbon” in CH₄ fluxes emitted from wetlands indicated a potential link between the gross primary productions (GPP) and CH₄ emissions (Chanton et al., 1995). Moreover, King and Reedburgh (2002) connected CH₄ production with the photosynthesis (a surrogate of biomass) through root exudation. Future research should investigate whether there is a clear mechanism (e.g., mediated by the primary productivity) to drive CH₄ productions and emissions.

The type, quality and availability of the substrate, plant type, and CH₄ emission pathways (ebullition, diffusion) also play an important role in the net emission from the salt marshes. Previous studies reported a wide range (30-100%) of contribution of plant-mediated CH₄ fluxes on overall CH₄ emission (Dorodnikov et al., 2011; Bridgham et al., 2013). However, this study focused on the relative control of the abiotic factors to mechanistically link and quantifies their contribution in regulating CH₄ emission fluxes.

7.4.2 Parsimonious predictive modeling

Most of the available predictive CH₄ emission models are process based and these models incorporated detail mechanisms (from microbial level to the emission pathways) of CH₄ production and emission. However, such detailing involved a large parameter set and a thorough understanding of the system. On the contrary, the model developed in the study only needs three input variables (soil temperature, water depth, and soil salinity) to predict the CH₄ emission from the wetlands. Bubier and Moore (1995) developed a predictive CH₄ emission model as a function

of water depth by using a weighted averaging regression technique for Canadian peatlands and the model estimated daily emissions with a good fitting efficiency (correlation coefficient, r^2 varied between 0.70 to 0.88 for different peatland sites). However, the absence of temperature and biogeochemical processes in the predictor set made the model exclusively site specific and hinders its application for different climate change scenarios. Huang et al. (1998) proposed a semi-empirical model of methane emission for flood paddy soil by making parameterization for CH_4 production and emission separately. The methane emission model predicted the observations with a high correlation coefficient ($r^2=0.87$). Both of the models are applicable for freshwater wetlands unlike the model developed in this study, which is explicitly trained for the saline water wetlands.

The developed 4-parameter model (one scaling factor, and three shape factors representing the effect of WD, ST, and SS, respectively) (Equation 2), therefore, empirically incorporated the primary environmental and hydrological factors associated with CH_4 emission, and the model robustness and parsimony were preserved because of the consideration of such small predictor set. Moreover, the quantitative estimation of the instantaneous fluxes from the small set of environmental variables leveraged the estimation of the annual scale ecosystem carbon balance (NECB), since CH_4 uptake is one of the most important components of NECB. An alternative model (Equation 3) was developed that only considers WD as input for the partially submerged condition ($\text{WD}<0$). However, it is strongly recommend to use the 4-parameter model (Equation 2), since it covers all the important process variables and the model could be helpful to estimate CH_4 emissions under different climate change (increase of temperature, sea level rise and change in salinity gradient) scenarios.

Since the final model was developed from the mean of the parameters estimated from 10000 iterations and parameters relatively followed the normal distribution with the significantly

small standard deviations (Figure 7.4, 7.5, and 7.6), the uncertainties of the parameters were less. Moreover, the modeling performance during calibration and validation also suggested relatively less uncertainty in the model. The final predictive model described 79% and 89% (Table 7.5) of the data variances for $WD > 0$ and $WD < 0$, respectively showing the elegance of data-driven empirical modeling in predicting such complex mechanism.

7.4.3 Significance of the estimated parameters

As the model was fitted and tested in the original untransformed domain, the importance of each of the input variable in the model cannot be explained based on the magnitude of the associated shape factors; this is not necessary because the importance and relative linkages of the variable were already investigated by the explanatory PLSR model. In this section, we tried to explain the significance of the estimated parameters on the basis of their signs.

The scale factor, a does not represent or explain the underlying shape of the physical parameters in the fitting. Rather a resembles the limit of the equation in the fitting domain. The reason behind to the estimation of the scaling factor as a power of 10 was that it eliminated the model limitation imposed by the scaling factor; If the scaling parameter is estimated without raising to the power of 10 while fitting the data, the value of a can be 0 for a particular combination set during iteration and that could estimate the F_{CH_4} as zero. By raising the a to the power of 10, this limiting factor was avoided.

For the $WD > 0$ condition the shape factors were estimated as 3.54, -0.19, and -1.17 for ST, WD, and SS, respectively (Table 7.5). The positive sign of the parameter of ST indicated a positive relation with the F_{CH_4} while the negative sign associated with WD and SS indicated a negative relation. Since the shape factor is greater than 1 for ST, the F_{CH_4} increases exponentially with the increase of soil temperature. A previous study also reported a similar phenomenon of

higher positive temperature dependency on CH₄ emission during completely submerged condition (Walter et al., 2000). However, it was quite difficult to readily describe the negative value of the shape factor associated with WD, since per our mechanistic knowledge CH₄ emission should increase with the increase of WD. There could be two possible explanations for the negative relation of WD. Firstly, water could act as a barrier to the diffusion of CH₄ from the soil when flooding condition persists predominantly. That could also refer a potential of higher CH₄ emission for the nearly submerged condition than the completely submerged condition (Cheng et al., 2007). The second possible explanation could be the use of all the available substrates for the CH₄ production if the water table was constantly above the soil surface. That might reduce the CH₄ emission rate in a substrate limiting environment. The negative value of the parameter corresponds to SS was consistent with our mechanistic understating. As more salt water intrudes in the marshes (an increase of flood depth), the CH₄ production and emission will be decreased as discussed previously.

However, for the WD<0 condition, the estimated shape factors (-0.95, 1.10, and -1.20 for ST, WD, and SS, respectively) (Equation 2) revealed different biogeochemical attributes. The positive sign of the parameter regarding WD referred a positive link between the WD and F_{CH₄}. However, the negative sign of the parameter associated with ST indicated a decrease of methane flux with an increase in soil temperature. In opposed to our general mechanistic knowledge of F_{CH₄}-temperature linkage, Hoehler and Alperin (2014) argued a possible negative effect on CH₄ emission due to temperature change, which was referred as a “piece of puzzle” in the climatic research. They provided an evidence of higher methane emission at low temperature as contrary to the well-known F_{CH₄}-temperature relationship described in Arrhenius equation. However, their study did not distinguish the flood and non-flood conditions. The different type of temperature dependency for the varying hydrologic condition that emerged from the study is subjected to much detail investigation. One possible explanation for such negative dependence could be the

higher rate of evapotranspiration at a higher temperature level that could decrease the CH₄ emission rate. Moreover, dominant positive linkage of the water table and CH₄ emission for the WD<0 condition might limit the temperature dependency (Roulet et al., 1993; Valentine et al., 1994). Similarly, the positive sign of the soil salinity parameter during the no flood condition could be due to the complete control of WD on CH₄ production. The emission rate could be higher when the system is nearly submerged because of the more convenient emission pathways than the submerged condition even though salinity is increasing. For the two-parameter model (CH₄ emission as a function of WD) (Equation 3), the estimated parameter value of WD was higher than one and illustrates the similar positive relation with F_{CH₄} as discussed before.

The estimated coefficients of the predictive model provided critical insights on how the most important process drivers are linked with the wetland methane emissions at varying hydrologic conditions. However, some of the relations that were revealed from this study are subjected to further investigations by using data from a range of coastal wetlands.

The developed CH₄ flux model (equation 2) was incorporated into an Excel spreadsheet along with the wetland CO₂ flux model to estimate the net ecosystem carbon balance of a given wetland site. The Excel-based model was named as “BWM Wetland C and GHG Model” and developed as a part of recently completed project titled “Bringing wetland to market: nitrogen and coastal blue carbon” funded by NOAA. The incorporation of the model in the Excel spreadsheet leverages the prediction of the GHG fluxes in a much easier way. The Excel-based model can be considered as a user-friendly engineering tool that required minimum input data. The BWM wetland C model is available in the Waquoit Bay National Estuarine Research Reserve website (<http://www.waquoitbayreserve.org/research-monitoring/salt-marsh-carbon-project/>), and detail information of the model and the user-guide is also available as a form of factsheet in the

Waquoit bay website (http://www.waquoitbayreserve.org/wp-content/uploads/BWM-Model-for-Predicting-Greenhouse-Gas-Fluxes_FS_FINAL.pdf)

7.5 Conclusions

The relative controls and linkages of different physical variables on CH₄ emission fluxes coastal salt marshes of four different wetlands of Waquoit Bay, MA were determined by using a robust multivariate data-analytics framework. The linkages were then leveraged to develop a parsimonious (minimum parameter), nonlinear, data-driven model for the prediction of the CH₄ emission of the study wetlands along an ecological gradient. The data-analytics revealed high control of temperature and hydrology (soil temperature and water table position) on CH₄ emissions during completely submerged condition (when water table is above the marsh surface). On the contrary, hydrology and soil biogeochemistry (water table position and soil salinity) demonstrated strong CH₄ flux linkages during partially submerged condition (when the water table is below the marsh surface). The empirical predictive model of instantaneous CH₄ fluxes was developed as a nonlinear function of the most three important drivers (water depth, soil temperature, and soil salinity) as determined from the data-analytics framework. The developed data-driven model could be used as a supplement to the existing process-based models to provide empirical foundation in obtaining critical decisions on “blue carbon” sequestration.

References

- Bridgham, S. D., Cadillo-Quiroz, H., Keller, J. K., Zhuang, Q., 2013. Methane emissions from wetlands: biogeochemical, microbial, and modeling perspectives from local to global scales. *Global Change Biology*, 19(5), 1325-1346.
- Bubier, J. L., Moore, T. R., 1994. An ecological perspective on methane emissions from northern wetlands. *Trends in ecology & evolution*, 9(12), 460-464.
- Bubier, J. L., Moore, T. R., Juggins, S., 1995. Predicting methane emission from bryophyte distribution in northern Canadian peatlands. *Ecology*, 677-693.
- Chanton, J. P., Bauer, J. E., Glaser, P. A., Siegel, D. I., Kelley, C. A., Tyler, S. C., ... Lazrus, A., 1995. Radiocarbon evidence for the substrates supporting methane formation within northern Minnesota peatlands. *Geochimica et Cosmochimica Acta*, 59(17), 3663-3668.
- Cheng, X., Peng, R., Chen, J., Luo, Y., Zhang, Q., An, S., ... Li, B., 2007. CH₄ and N₂O emissions from *Spartina alterniflora* and *Phragmites australis* in experimental mesocosms. *Chemosphere*, 68(3), 420-427.
- Crooks, S., Emmett-Mattox, S., Findsen, J., 2010. Findings of the National Blue Ribbon Panel on the development of a greenhouse gas offset protocol for tidal wetlands restoration and management: action plan to guide protocol development. Restore America's Estuaries, Philip Williams & Associates, Ltd., and Science Applications International Corporation.
- Crooks, S., Herr, D., Tاملander, J., Laffoley, D., Vandever, J., 2011. Mitigating climate change through restoration and management of coastal wetlands and near-shore marine ecosystems: challenges and opportunities.
- Cui, J., Li, C., Sun, G., Trettin, C., 2005. Linkage of MIKE SHE to Wetland-DNDC for carbon budgeting and anaerobic biogeochemistry simulation. *Biogeochemistry*, 72(2), 147-167.
- Dorodnikov, M., Knorr, K. H., Kuzyakov, Y., Wilmking, M., 2011. Plant-mediated CH₄ transport and contribution of photosynthates to methanogenesis at a boreal mire: a ¹⁴C pulse-labeling study. *Biogeosciences*, 8(8), 2365-2375.
- Dunfield, P., Dumont, R., Moore, T. R., 1993. Methane production and consumption in temperate and subarctic peat soils: response to temperature and pH. *Soil Biology and Biochemistry*, 25(3), 321-326.
- Garnet, K. N., Megonigal, J. P., Litchfield, C., Taylor, G. E., 2005. Physiological control of leaf methane emission from wetland plants. *Aquatic Botany*, 81(2), 141-155.
- Hartmann, D.L.; Klein Tank, A.M.G.; Ruscicucci, M.; Alexander, L.V.; Broenniman, B.; Charabi, Y.; Dentener, F.J.; Dlugokencky, E.J.; Easterling, D.R.; Kaplan, A.; Soden, B.J.; Thorne, P.W.; Wild, M.; Zhai, P.M.; Kent, E., 2013 Observations: Atmosphere and Surface. In: Stocker, T.F.; Qin, D.; Plattner, G.-K.; Tignor, M.; Allen, S.K.; Boschung, J.; Nauels, A.; Xia, Y.; Bex, V.; Midgley, P.M., (eds.) *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, Cambridge University Press, 159-254.

- Hoehler, T. M., Alperin, M. J., 2014. Biogeochemistry: Methane minimalism. *Nature*, 507(7493), 436-437.
- Howarth, R. W., 2014. A bridge to nowhere: methane emissions and the greenhouse gas footprint of natural gas. *Energy Science & Engineering*, 2(2), 47-60.
- Huang, Y., Sass, R. L., Fisher Jr, F. M., 1998. A semi-empirical model of methane emission from flooded rice paddy soils. *Global Change Biology*, 4(3), 247-268.
- King, J. Y., Reeburgh, W. S., 2002. A pulse-labeling experiment to determine the contribution of recent plant photosynthates to net methane emission in arctic wet sedge tundra. *Soil Biology and Biochemistry*, 34(2), 173-180.
- Megonigal, J. P., Mines, M. E., Visscher, P. T., 2004. Anaerobic metabolism: linkages to trace gases and aerobic processes. In: Schlesinger WH (ed) *Biogeochemistry*, Elsevier-Pergamon, Oxford, pp 317-424.
- Moore, T. R., De Young, A., Bubier, J. L., Humphreys, E. R., Lafleur, P. M., Roulet, N. T., 2011. A multi-year record of methane flux at the Mer Bleue Bog, Southern Canada. *Ecosystems*, 14(4), 646-657.
- Morris, J. T., Edwards, J., Crooks, S., Reyes, E., 2012. Assessment of carbon sequestration potential in coastal wetlands. In *Recarbonization of the Biosphere* (pp. 517-531). Springer Netherlands.
- Nellemann, C., Corcoran, E., (Eds.) 2009. *Blue carbon: the role of healthy oceans in binding carbon: a rapid response assessment*. UNEP/Earthprint.
- Poffenbarger, H. J., Needelman, B. A., Megonigal, J. P., (2011). Salinity influence on methane emissions from tidal marshes. *Wetlands*, 31(5), 831-842.
- Roulet, N. T., Ash, R., & Moore, T. R., 1992. Low boreal wetlands as a source of atmospheric methane. *Journal of Geophysical Research: Atmospheres* (1984–2012), 97(D4), 3739-3749.
- Roulet, N. T., Ash, R., Quinton, W., Moore, T., 1993. Methane flux from drained northern peatlands: effect of a persistent water table lowering on flux. *Global Biogeochemical Cycles*, 7(4), 749-769.
- Segers, R., 1998. Methane production and methane consumption: a review of processes underlying wetland methane fluxes. *Biogeochemistry*, 41(1), 23-51.
- Stocker, T. F., Qin, D., Plattner, G. K., Tignor, M., Allen, S. K., Boschung, J., ... Midgley, B. M., 2013. IPCC, 2013: climate change 2013: the physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change.
- Valentine, D. W., Holland, E. A., Schimel, D. S., 1994. Ecosystem and physiological controls over methane production in northern wetlands. *Journal of Geophysical Research: Atmospheres* (1984–2012), 99(D1), 1563-1571.

- Walter, B. P., Heimann, M., 2000. A process-based, climate-sensitive model to derive methane emissions from natural wetlands: Application to five wetland sites, sensitivity to model parameters, and climate. *Global Biogeochemical Cycles*, 14(3), 745-765.
- Waquoit bay national estuarine research reserve website, 2015
<http://www.waquoitbayreserve.org/>
- Yvon-Durocher, G., Allen, A. P., Bastviken, D., Conrad, R., Gudas, C., St-Pierre, A., ... Del Giorgio, P. A., 2014. Methane fluxes show consistent temperature dependence across microbial to ecosystem scales. *Nature*, 507(7493), 488-491.
- Zhu, G. B., Zhou, L. L., Jiang, B., Wang, Y., Wang, S. Y., Jetten, M. S. M., Yin, C. Q., 2014. Biogeographical distribution of nitrite dependent anaerobic methanotrophic bacteria in wetland ecosystems: from regional scale to global scale. NCBI Genbank web <http://www.ncbi.nlm.nih.gov/nuccore/>.

Chapter 8: Investigation of biogeochemical similitudes, environmental regimes, and scaling relationships of greenhouse gas (GHG) fluxes from coastal wetlands

Abstract

Wetland is one of the complex geographic features in the earth and subjected to an intricate process of biotic-abiotic interactions and the exchanges of greenhouse gasses (GHGs) of CO₂ and CH₄. Based on the analysis of the observed data of four coastal salt marsh wetlands of Waquoit Bay, MA, the study explored the potential of finding similitudes, environmental regimes, and scaling relationships of GHG fluxes from coastal wetlands. The study identified two important dimensionless physico-chemical groups to predict the dimensionless GHG flux groups. The study revealed the existence of two different environmental regimes governed by the hydrological and biogeochemical process variables; the emergent environmental regimes showed that the coastal wetland system becomes most productive (high sequestration and low emission) when the water table position is at the surface. Different power-law based dimensionless scaling relations were obtained for flood and non-flood conditions for the dimensionless CO₂ and CH₄ groups. The scaling relations were later expressed in a dimensional domain to find the GHG fluxes scaling relations with the physical drivers. The power-law based scaling of PAR was identified as the most dominating factor in CO₂ sequestration whereas the position of the water table was the most influential scaling factor for the CH₄ emission. The research provided a crucial knowledge into the wetland biogeochemical emergence and similitudes that can lead to the development of the parsimonious (minimum parameter), spatiotemporally robust predictive models of GHG emissions and carbon sequestration under a current change of climate and sea level rise.

8.1 Introduction

Wetlands play a critical role in soil-atmospheric exchanges of the major GHGs of carbon dioxide (CO_2) and methane (CH_4). On a molar basis CH_4 yield 34 and 86 times the global warming potential of CO_2 for 100 years and 20 years, respectively (Stocker et al., 2013, Howarth 2014). Because of its unique biogeochemical characteristics, the carbon (C) sequestration rate of the tidal marshes (i.e., coastal wetlands) is ranked first among the different ecosystems (Chmura et al., 2003, Duarte et al., 2005, and Bridgham et al., 2007). Therefore, tidal wetlands have the potential to play a pivotal role in global warming mitigation. However, an overarching science question is how the C storage and GHG flux rates of the coastal wetlands respond to the changes in climate (e.g., temperature, precipitation), sea level rise (SLR) and inundation (e.g., soil moisture, salinity), and land managements (e.g., nutrient loading)? What are the different environmental regimes of wetland GHG emissions and C sequestration? Can we identify any biogeochemical similarities for GHG emissions and sequestered C? Do the emissions and sequestrations scale in time and space?

Fluxes of GHGs are produced through soil microbial processes under the influence of climate and soil physicochemical factors (Smith et al., 2003). Conrad (1989) described wetland methane (CH_4) emissions as a result of the following soil biogeochemical processes: (i) CH_4 production by methanogenic bacteria under anaerobic conditions; (ii) CH_4 oxidation by methanotrophic bacteria mainly under aerobic conditions; and (iii) CH_4 transport to the atmosphere. Major predictors of CH_4 emissions are water table position (determining the anoxic soil zone for CH_4 production), soil temperature (influencing the rates of organic matter degradation, CH_4 production and oxidation), vegetation, soil carbon content, pH, salinity, sulfate, and pathways of CH_4 transportation to the atmosphere (Conrad 1989, Wang et al., 1996, Walter and Heimann 2000). Previous research (e.g., Sebacher et al., 1986) reported a positive correlation

of CH₄ emissions to the atmosphere with soil temperature and moisture content. Bartlett et al. (1987) and Poffenbarger et al. (2011) reported a strong negative correlation between CH₄ emissions and soil salinity over a gradient from freshwater to highly saline (> 26 ppt) tidal marshes. Wetland CO₂ sequestration fluxes are mainly driven by vegetation biomass and light (i.e., photosynthesis), temperature, salinity, pH, water table position, moisture content, and sulfate (Smith et al., 2003).

The influence of these factors on the biological processes (CO₂ sequestration and CH₄ emission) is not independent. Rather all the factors are connected to each other and collectively drive the GHG exchanges. The linkage between the process factors and GHG fluxes are highly complex and nonlinear. Yu (2006) reported that peatland environmental and biological variables follow a power-law function at different temporal scales. Moreover, GHG fluxes varied spatially depending on the salinity, nutrient, and vegetation gradient. The spatiotemporal variability of the different factors that controls the coastal wetland flux exchanges hinders the development of a generalized scaling relations and modeling framework across wetland salt marshes.

A crucial advantage of identifying similitude and dimensionless parameters is that estimations and predictions for other systems can be obtained from equations and curves developed from a few sites representing different gradients of response and predictor variables. In classical fluid mechanics and hydraulic engineering, similitude is defined as the parametric reduction of physical problems to develop independent dimensionless numbers and a dimensionless formulation of the physical system (Kundu and Cohen 2004). Similitude can be obtained through dimensional analysis using Buckingham pi-theorem (particularly useful for systems lacking sound and simple governing equations) and/or through normalizing (i.e., scaling) the underlying governing equations by appropriate characteristic parameters (e.g., length, velocity, density).

A classical hydraulic engineering example of similitude and scaling applications can be the Moody chart for pipe flow design. It is a dimensionless scaling formulation applicable to different combinations of flow velocity, viscosity, density, as well as pipe surface roughness, diameter and length for determining dimensionless friction factors based on dimensionless Reynolds Number (Re) and relative roughness (dimensionless ratio of pipe surface roughness to pipe diameter, ϵ/d). The friction factor quantifies the friction head loss (i.e., pressure drop) along the pipe length. The reduction of 7 parameters for pipe flow design (flow pressure, velocity, viscosity, density, pipe roughness, diameter, and length) to 3 dimensionless parameters (friction factor, Re , and ϵ/d) is called “similitude” and the functional relationships between them are “scaling relationships” that can have different forms depending on the flow regimes (laminar, transitional, and turbulent), as defined by the critical values of Re . Abdul-Aziz et al. (2007a, b) presented an important scaling application in aquatic health and ecological engineering. They conceptually identified a site-specific, reference-time (e.g., noon) dissolved oxygen (DO) measurement as a scaling parameter to normalize hourly DO data from different streams and obtain a general, dimensionless diurnal cycle. The dimensionless formulation resulted in robust predictions of DO both in time (hours, days, and seasons) and space (different streams representing various distinct ecoregions of Minnesota).

Chapter 2 provides an important evidence of ecological similitude in terrestrial Net Ecosystem Exchange (NEE). Different diurnal cycles representing different days of NEE of five deciduous forests were scaled to a dimensionless curve from where the hourly NEE were predicted. The estimated parameters were spatiotemporally robust and did not show any sensitivity with the major driver of the NEE. The means of the parameters represent a generalized modeling parameter set for the prediction of NEE of deciduous forest. Warnars et al. (2007) applied dimensional analysis to identify similitude in stream biogeochemistry and ecology and developed useful scaling relationships between biotic (e.g., biomass) and abiotic (e.g., climatic,

hydrologic, and geomorphic) variables in different streams and rivers across North America. West et al. (2001) demonstrated similitude and robust scaling in biology by deriving a single, parameter less universal curve to describe the growth rate of many diverse species, including shrimp, salmon, heron, cow, etc.

The goal of this study is to unravel wetland similitudes, environmental regimes in wetland greenhouse gas (GHG) emissions and carbon sequestration and apply them to formulate spatiotemporally robust scaling relationships. The research tests a fundamental hypothesis that GHG emissions and C sequestration from wetland ecosystems follow distinct biogeochemical similitudes and robust scaling relationships. The dimensional analysis is applied to formulate useful dimensionless numbers, identify similitude and environmental regimes, and scaling relationships for major GHG (CO_2 , CH_4) fluxes from wetland ecosystems. The similitude and scaling relationships were evaluated from the observed data of four different coastal wetland ponds of Waquoit Bay, MA. The research is expected to provide crucial knowledge into the wetland biogeochemical emergence pattern and to identify different environmental regimes and the associated transitional threshold for the robust development of the predictive model of wetland GHG fluxes.

8.2 Materials and methods

8.2.1 Study sites and data set

The same study wetland ponds (Sage lot pond, Eel pond, great pond and Hamblin pond) that were used for the development of the daytime CO_2 exchange and CH_4 emission models, were also used in this study (please see chapter 6 and 7, section 6.2.1) (Waquoit Bay website, 2015)

Instantaneous greenhouse gas (GHG) fluxes (CO_2 and CH_4) and corresponding biogeochemical, hydrological and climatic variables were measured in different days between

June to November, 2013 (5 days in June, 2 days in July, 4 days in August, 2 days in September, 2 days in October, and 2 days in November) in the low marsh zone of the study ponds by using the state-of-the-art technologies. The complete dataset included GHG fluxes and five hydrological, biogeochemical and climatic variables (Table 8.1): CO₂ sequestration flux (F_{CO2}), CH₄ emission flux (F_{CH4}), photosynthetic active radiation (PAR), soil temperature (ST), soil salinity (pore water) (SS), pH, and water depth (WD; water level is converted to water depth relative to the marsh surface). The predictor variables were selected based on the vigorous analysis conducted in the dimensional domain for the identification of the dominant hydro-climatic and biogeochemical predictors responsible for CO₂ uptake and CH₄ emissions. The air temperature was not considered in this analysis since it was used to calculate the GHG fluxes. Instead as a surrogate of the ambient climate, soil temperature was included in the analysis. The measurement methodology of the fluxes and other participatory variables were discussed in chapter 6 (see section 6.2.1).

Table 8.1: List of variables, units, and data range and mean for all the study wetlands.

Variable	Unit	Range	Mean
CO ₂ sequestration Flux (F _{CO2})	μmol/m ² /s	-0.05 to -19.62	-7.34
CH ₄ emission Flux (F _{CH4})	μmol/m ² /s	0.000096 to 0.0077	0.0016
Soil Temperature (ST)	°C	8.89 to 27.83	20.19
Sediment Salinity (SS)	ppt	20 to 36	29.58
Photosynthetic Active Radiation (PAR)	μmol/m ² /s	115.56 to 2080.67	1424.18
Water Depth (WD)	m	-0.075 to 0.329	0.088
pH		3.83-7.80	6.804

8.2.2 Data preparation and data statistics

The data preparation, filtering and data statistics of F_{CO2} and F_{CH4} and corresponding environmental and biogeochemical variables were discussed in chapter 6 and 7 (Please see section 6.2.2 and 7.2.2)

8.2.3 Analysis methodology

The methodology was designed to achieve the study goals. Four-step analysis procedures were used to study the wetland ecological similitude and scaling relationships for F_{CO_2} and F_{CH_4} : (i) formulation of different dimensionless numbers through dimensional analysis, (ii) identification of important and meaningful dimensionless numbers from the observed data by using a multivariate data-analytics framework, (iii) examining similitude and scaling relationships across different environmental regimes for the F_{CO_2} and F_{CH_4} from the important dimensionless numbers, and (iv) developing scaling relations of CO_2 sequestration and CH_4 emission fluxes through non-linear modeling.

Formulation of different dimensionless numbers through dimensional analysis

Dimensional analysis, extensively used in fluid mechanics and hydraulic engineering (Kundu and Cohen 2004), was used to formulate dimensionless functional groups (or numbers) and derive scaling relationships between the dimensionless response variables (GHG emissions and sequestered carbon) and the independent predictor variables (hydro-climatic, and biogeochemical). A dimensional analysis involves different combinations of the predictor and response variables resulting in several dimensionless numbers or groups, which represent biogeochemical similitude and emergence patterns among different wetland ecosystems. Only the independent dimensionless groups/numbers were considered as predictor variables in developing a scaling relationship with the respective dimensionless response numbers.

The dimensions, units and notations of the variables considered in the dimensional analysis were given in Table 8.2. I considered PAR, ST, SS, pH, and WD as the predictor variables to drive the dimensionless groups for both F_{CO_2} and F_{CH_4} , which holds four fundamental dimensions (M, L, T, K). Since temperature dimension (K) is only available for ST (Table 8.2), I included specific heat of soil in the data matrix that has temperature dimension to balance the

temperature dimension. Moreover, time (T) was also considered as a variable in the dimensional analysis by hypothesizing that time would leverage the temporal scale transitions (i.e., instantaneous scale to hourly, daily, monthly scale) in future studies. Finally, eight variables that include the response variable (either F_{CO_2} or F_{CH_4}) were considered in the dimensional analysis. The most appropriate and meaningful set of dimensionless π numbers were obtained through a diligent iterations procedure and each iteration produced 4 π groups (number of total variables (8) – total fundamental dimensions (4)) that incorporated both predictor and response numbers.

Table 8.2: List of variables used for the dimensional analysis of wetland GHG fluxes.

Variable	Unit	Dimension	Notation
GHG Fluxes (F_{CO_2} and F_{CH_4})	$g/m^2/s$	$[M/L^2/T]$	A
Soil Temperature (ST)	Kelvin (K)	[K]	B
Sediment Salinity (SS)	g/m^3	$[M/L^3]$	C
Photosynthetic Active Radiation (PAR)	g/s^3	$[M/T^3]$	D
Specific Heat of Air (C_p)	J/g/K	$[L^2/T^2/K]$	E
Water Depth (WD)	m	[L]	F
Time (t)	s	[T]	G
Concentration of H^+	g/m^3	$[M/L^3]$	H

Identification of important dimensionless numbers from the observed data by using a multivariate data-analytics framework

Important and meaningful dimensionless numbers were identified separately for both CO_2 and CH_4 fluxes from the different combinations of numbers by using a modified form of the multivariate data-analytic framework proposed by Ishtiaq and Abdul-Aziz (2015). The data-analytics method included Pearson correlation matrix, explanatory principal component analysis (PCA), and factor analysis (FA). Results obtained from PCA were presented through biplots by considering 1st two major principal components (PCs). The optimal number of factors were extracted based on an initial eigenvalue criterion (eigenvalue>1) to ensure that the extracted factors optimally explained the overall data variance. Moreover, the varimax orthogonal rotation was performed in the FA to maximize the loadings (i.e., a correlation between a factor and a

variable) of the variables on each extracted factor. Since different variables were in different unitary systems, they were converted to SI units by using appropriate conversion factors (Table 8.3).

Table 8.3: Unit conversion factors of all the variables used in dimensional analysis

Variables	Original Unit (X)	Converted Unit (Y)	Conversion
F _{CO2}	micromole/m ² /S	gC/m ² /s	Y = 0.000012 * X
F _{CH4}	micromole/m ² /S	gC/m ² /s	Y = 0.000012 * X
ST	°C	K	Y = 273.15 + X
C _p (air) = 1.005	kJ/kg/K	J/g/K	Y = X
PAR	micromole/m ² /s	g/s ³	Y = 212.77 * X
SS	ppt	g/m ³	Y = 1000 * X
H ⁺	g/L	g/m ³	Y = 1000 * X
WD	m	m	Y=X
t	s	s	Y=X

Note: F_{CO2} and F_{CH4} conversion: Micromole is converted to mole and then multiplied by the molecular weight of C (M=12 g); PAR: micromole/m²/s was converted to W/m² (kg/s³) by applying a multiplication factor of 0.21277 and then multiplied by 1000 to convert kg/s³ to g/s³ (McCree 1981); ppt conversion: 1 ppt= 1000 mg/L=1000 g/m³.H⁺: 10^{-pH}*2.02* 1000 g/m³, where 2.02 is the molecular weight of hydrogen (H₂).

Examining similitude and scaling relationships across different environmental regimes for the F_{CO2} and F_{CH4} from the important dimensionless numbers

At first, the most important and meaningful predictor dimensional numbers that were identified through the data-analytic framework were plotted against the response dimensionless number (for both F_{CO2} and F_{CH4}) for different spatial (i.e., across sites) conditions. Flood and dry conditions (based on WD>0 and WD<0) were considered as the hydro-climatic and environmental regimes to examine the existence of similitude and scaling relationship, if any, in the derived groups. A potential collapse of the different groups in a comparable zone (Shield type diagram) would indicate the existence of ecological similitude and scaling relationships in GHG fluxes.

Secondly, the nonlinear scaling relationships ($Y = X^b$) among the predictor (X) and response (Y) dimensionless groups were investigated by applying the Levenberg-Marquardt non-

linear least squares algorithm. The comparable values of the shape factor (b) in different regimes would indicate the potential similitude and scaling relations among the response and predictor groups that could lead to the robust predictive model development in the dimensionless domain.

Developing scaling laws of CO₂ sequestration and CH₄ emission fluxes through non-linear modeling

Wetland similitudes and scaling laws can be utilized to formulate empirical models for robust predictions of GHG fluxes. The choice of the empirical equation (e.g., power law, polynomial, and linear) depends on the functional relationship between the response (Y) and predictor (X) dimensionless groups. The power law based predictive empirical models (

$Y = 10^b X_1^{b_1} X_2^{b_2} \dots X_n^{b_n}$ type) was used for the model development. Ideally, the model parameters (b, b₁...b_n) should be same across all the environmental regimes; comparable parameter values in a range of spatial and environmental gradient would develop the spatially robust predictive models of CO₂ sequestration and CH₄ emission.

8.3 Results

8.3.1 Formulation of dimensionless numbers

According to the Buckingham pi theorem, different combinations of repetitive primary variables can be chosen to form various pi groups. Different iterations were performed while formulating the pi groups by changing the primary variables. As examples, the detailed derivations of the two possible alternatives were given below, which provided most significant dimensionless groups:

Iteration 1

Number of Variables, n = 8

Considered Dimensions, m = 4

Number of dimensionless groups = n-m = 4

$$F(A, B, C, D, E, F, G, H) = 0$$

$$\phi(\Pi_1, \Pi_2, \Pi_3, \Pi_4) = 0$$

Considering B, C, F and G as repeating primary variables, possible Π terms should be

$$\Pi_1 = B^a C^b F^c G^d A$$

$$\Pi_2 = B^a C^b F^c G^d D$$

$$\Pi_3 = B^a C^b F^c G^d E$$

$$\Pi_4 = B^a C^b F^c G^d H$$

Π_1 :

$$M^0 L^0 T^0 K^0 = (K)^a \left(\frac{M}{L^3}\right)^b (L)^c (T)^d \left(\frac{M}{L^2 T}\right)$$

$$K : a = 0$$

$$M : b = -1$$

$$L : -3b + c - 2 = 0; c + 3 = 2; c = -1$$

$$T : d = 1$$

$$\Pi_1 = (B)^0 (C)^{-1} (F)^{-1} (G)^1 (A)$$

$$\Pi_1 = \frac{AG}{CF}$$

Π_2 :

$$M^0 L^0 T^0 K^0 = (K)^a \left(\frac{M}{L^3}\right)^b (L)^c (T)^d \left(\frac{M}{T^3}\right)$$

$$K : a = 0$$

$$M : b = -1$$

$$L : -3b + c = 0; c = -3$$

$$T : d - 3 = 0; d = 3$$

$$\Pi_2 = (B)^0 (C)^{-1} (F)^{-3} (G)^3 (D)$$

$$\Pi_2 = \frac{DG^3}{CF^3}$$

Π_3 :

$$M^0 L^0 T^0 K^0 = (K)^a \left(\frac{M}{L^3}\right)^b (L)^c (T)^d \left(\frac{L^2}{T^2 K}\right)$$

$$K : a = 1$$

$$M : b = 0$$

$$L : -3b + c + 2 = 0; c = -2$$

$$T : d = 2$$

$$\Pi_3 = (B)^1 (C)^0 (F)^{-2} (G)^2 (E)$$

$$\Pi_3 = \frac{BEG^2}{F^2}$$

$$\Pi_4 :$$

$$M^0 L^0 T^0 K^0 = (K)^a \left(\frac{M}{L^3}\right)^b (L)^c (T)^d \left(\frac{M}{L^3}\right)$$

$$K : a = 0$$

$$M : b = -1$$

$$L : -3b + c - 3 = 0; c = 0$$

$$T : d = 0$$

$$\Pi_4 = (B)^0 (C)^{-1} (F)^0 (G)^0 (H)$$

$$\Pi_4 = \frac{C}{H}$$

$$\phi \left[\left(\frac{F_{CO2/CH4} * t}{WD * SS} \right) \left(\frac{PAR * t^3}{WD^3 * SS} \right) \left(\frac{ST * C_p * t^2}{WD^2} \right) \left(\frac{SS}{H^+} \right) \right]$$

Iteration 2

Number of Variables, n = 8

Considered Dimensions, m = 4

Number of dimensionless groups = n-m = 4

$$F(A, B, C, D, E, F, G, H) = 0$$

$$\phi(\Pi_1, \Pi_2, \Pi_3, \Pi_4) = 0$$

Considering B, D, F and G as repeating primary variables, possible Π terms should be

$$\Pi_1 = B^a D^b F^c G^d A$$

$$\Pi_2 = B^a D^b F^c G^d C$$

$$\Pi_3 = B^a D^b F^c G^d E$$

$$\Pi_4 = B^a D^b F^c G^d H$$

Π_1 :

$$M^0 L^0 T^0 K^0 = (K)^a \left(\frac{M}{T^3}\right)^b (L)^c (T)^d \left(\frac{M}{L^2 T}\right)$$

$$K : a = 0$$

$$M : b = -1$$

$$L : c = 2$$

$$T : -3b + d - 1 = 0; d = -2$$

$$\Pi_1 = (B)^0 (D)^{-1} (F)^2 (G)^{-2} (A)$$

$$\Pi_1 = \frac{AF^2}{DG^2}$$

Π_2 :

$$M^0 L^0 T^0 K^0 = (K)^a \left(\frac{M}{T^3}\right)^b (L)^c (T)^d \left(\frac{M}{L^3}\right)$$

$$K : a = 0$$

$$M : b = -1$$

$$L : c = 3$$

$$T : -3b + d = 0; d = -3$$

$$\Pi_2 = (B)^0 (D)^{-1} (F)^3 (G)^{-3} (C)$$

$$\Pi_2 = \frac{DG^3}{CF^3}$$

Π_3 :

$$M^0 L^0 T^0 K^0 = (K)^a \left(\frac{M}{T^3}\right)^b (L)^c (T)^d \left(\frac{L^2}{T^2 K}\right)$$

$$K : a = 1$$

$$M : b = 0$$

$$L : c = -2$$

$$T : -3b + d - 2 = 0; d = 2$$

$$\Pi_3 = (B)^1 (D)^0 (F)^{-2} (G)^2 (E)$$

$$\Pi_3 = \frac{BEG^2}{F^2}$$

$$\Pi_4 :$$

$$M^0 L^0 T^0 K^0 = (K)^a \left(\frac{M}{T^3}\right)^b (L)^c (T)^d \left(\frac{M}{L^3}\right)$$

$$K : a = 0$$

$$M : b = -1$$

$$L : c = 3$$

$$T : -3b + d = 0; d = -3$$

$$\Pi_4 = (B)^0 (D)^{-1} (F)^3 (G)^{-3} (H)$$

$$\Pi_4 = \frac{DG^3}{HF^3}$$

$$\phi \left[\left(\frac{F_{CO2/CH4} * WD^2}{PAR * t} \right) \left(\frac{PAR * t^3}{WD^3 * SS} \right) \left(\frac{ST * C_p * t^2}{WD^2} \right) \left(\frac{PAR * t^3}{WD^3 * H^+} \right) \right]$$

The dimensional analysis generated two dimensionless response groups and four predictor dimensionless groups. For simplicity in presenting the dimensionless groups, I denoted the derived dimensionless groups with R (response group) and P (predictor group) as follows:

$$R_1 = \frac{F_{CO2/CH4} * t}{WD * SS} \text{ (dimensionless response group that includes } F_{CO2/FCH4}\text{)}$$

$$R_2 = \frac{F_{CO2/CH4} * WD^2}{PAR * t} \text{ (dimensionless response group that includes } F_{CO2/FCH4}\text{)}$$

$$P_1 = \frac{PAR * t^3}{WD^3 * SS} \text{ (radiation-redox group)}$$

$$P_2 = \frac{ST * C_p * t^2}{WD^2} \text{ (hydro-climate group)}$$

$$P_3 = \frac{SS}{H^+} \text{ (salinity-redox group)}$$

$$P_4 = \frac{PAR * t^3}{WD^3 * H^+} \text{ (radiation-redox group)}$$

8.3.2 Identification of important and meaningful dimensionless numbers

CO₂ sequestration flux

Pearson correlation analysis

Among the response groups, the R₁ showed high correlation with P₁, P₂, and P₄ (Table 8.4).

Since P₁ and P₄ were relatively similar in terms of mechanism, it can be stated that the “radiation-redox” and “hydro-climatic” groups were strongly linked with R₁. The other response group R₂ did not exhibit high correlation with the predictor groups; only weak to moderate correlation was observed between R₂ and P₃. Moreover, strong correlations were observed between P₁ and P₂, and P₂ and P₄; indicating the presence of a high level of multi-collinearity in the data matrix.

Principal component analysis

The first two PCs explained around 88% of the total data and system variance (Figure 8.1). Non-orthogonal (0 degree) orientations between the predictor groups (P₁, P₂, P₄) and response number, R₁ suggested high linkages. The P₃ and R₂ were orthogonally placed related to the P₁, P₂, P₄, and R₁ groups, indicating a very weak linkage between them. Overall, PCA analysis supported the observations obtained from the correlation analysis.

Table 8.4: Pearson correlation matrix for wetland CO₂ sequestration flux dataset for all the study ponds showing linear correlations among the participatory dimensionless numbers.

	P ₁	P ₂	P ₃	P ₄	R ₁	R ₂
P ₁	1.00					
P ₂	0.93	1.00				
P ₃	-0.07	-0.10	1.00			
P ₄	0.99	0.94	-0.03	1.00		
R ₁	0.87	0.93	-0.06	0.91	1.00	
R ₂	-0.11	-0.23	0.43	-0.13	-0.23	1.00

Notes: $R_1 = [(F_{CO_2} * t) / (WD * SS)]$; $R_2 = [(F_{CO_2} * WD^2) / (PAR * t)]$; $P_1 = [(PAR * t^3) / (WD^3 * SS)]$; $P_2 = [(ST * C_p * t^2) / WD^2]$; $P_3 = [SS / H^+]$; and $P_4 = [(PAR * t^3) / (WD^3 * H^+)]$. F_{CO_2} , PAR, WD, ST, SS, C_p , H^+ , and t refer, respectively, to net CO₂ sequestration flux, photosynthetic active radiation, water depth relative to soil surface, soil temperature, soil salinity, specific heat at constant pressure, hydrogen ion concentration, and time.

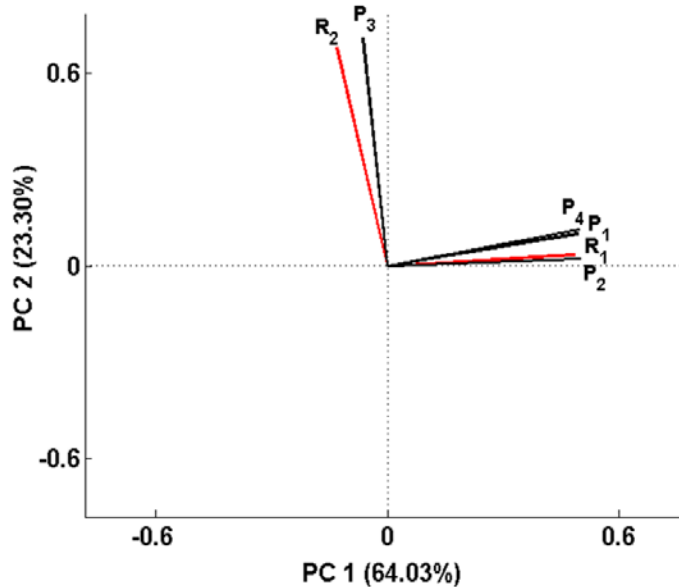


Figure 8.1: PCA biplots showing correlation strength and orientation of variables for wetland F_{CO_2} for all the study ponds of Waquoit Bay, MA. Percent variance explained by each PC is shown in parentheses. R_1 and R_2 are carbon flux group ($R_1 = [(F_{CO_2} * t) / (WD * SS)]$) and $R_2 = [(F_{CO_2} * WD^2) / (PAR * t)]$; P_1 and P_4 are the radiation-redox group ($P_1 = [(PAR * t^3) / (WD^3 * SS)]$) and $P_4 = [(PAR * t^3) / (WD^3 * H^+)]$; P_2 is the hydro-climatic group $[(ST * C_p * t^2) / WD^2]$; and P_3 is the salinity-redox group $[SS / H^+]$. F_{CO_2} , PAR, WD, ST, SS, C_p , H^+ , and t refer, respectively, to net CO₂ sequestration flux, photosynthetic active radiation, water depth relative to soil surface, soil temperature, soil salinity, specific heat at constant pressure, hydrogen ion concentration, and time.

Factor analysis

First two factors were extracted based on the initial eigenvalue criterion, eigenvalue >1 (Table 8.5). The high loadings of R₁, P₁, P₂, and P₄ on the first factor suggested their powerful linkages. None of the dimensionless groups loaded markedly on the factor 2 implying other response group R₂ did not have a genuine predictive ability.

The data-analytics showed that the “radiation-redox, P₁” and “hydro-climatic, P₂” numbers were the most meaningful and important number in explaining the response number, R₁ and R₁ can be expressed as a function of P₁ and P₂ (Equation 1). Other response group R₂ did not have the adequate predictive capability to be used for the further relationship analysis.

$$\frac{F_{CO_2} * t}{SS * WD} = f \left[\left(\frac{PAR * t^3}{WD^3 * SS} \right) \left(\frac{ST * C_p * t^2}{WD^2} \right) \right] \quad (1)$$

Table 8.5: Latent factors extracted from the data matrix for F_{CO₂}. Bold indicates variables that loaded highly on each factor.

Dimensionless group	P ₁	P ₂	P ₃	P ₄	R ₁	R ₂
Factor 1	0.99	0.96	-0.06	0.99	0.93	-0.14
Factor 2	-0.12	0.12	0.00	-0.04	0.37	-0.27

Notes: R₁=[(F_{CO₂}*t)/(WD*SS)]; R₂=[(F_{CO₂}*WD²)/(PAR*t)]; P₁=[(PAR*t³)/(WD³*SS)]; P₂ = [(ST*C_p*t²)/WD²]; P₃ =[SS/H⁺]; and P₄=[(PAR*t³)/(WD³*H⁺)]. F_{CO₂}, PAR, WD, ST, SS, C_p, H⁺, and t refer, respectively, to net CH₂ sequestration flux, photosynthetic active radiation, water depth relative to soil surface, soil temperature, soil salinity, specific heat at constant pressure, hydrogen ion concentration, and time.

CH₄ emission flux

Pearson correlation analysis

Among the response groups, the R₁ showed high correlation with P₁, P₂, and P₄ (Table 8.6). The other response group R₂ exhibited moderate correlations with P₁, P₃, and P₄; suggesting both R₁ and R₂ could potentially be used while forming scaling relations. P₃ did not show any notable correlation with the response groups. Moreover, strong correlations were observed between P₁

and P₂, P₁ and P₄, and P₂ and P₄; indicating the presence of a high level of multi-collinearity in the data matrix.

Table 8.6: Pearson correlation matrix for wetland CH₄ flux dataset for all the study ponds showing linear correlations among the participatory dimensionless numbers.

	P ₁	P ₂	P ₃	P ₄	R ₁	R ₂
P ₁	1.00					
P ₂	0.88	1.00				
P ₃	-0.07	-0.07	1.00			
P ₄	0.72	0.68	0.17	1.00		
R ₁	0.79	0.77	-0.16	0.61	1.00	
R ₂	-0.35	-0.43	0.18	-0.30	-0.41	1.00

$R_1 = [(F_{CH_4} * t) / (WD * SS)]$; $R_2 = [(F_{CH_4} * WD^2) / (PAR * t)]$; $P_1 = [(PAR * t^3) / (WD^3 * SS)]$; $P_2 = [(ST * C_p * t^2) / (WD^2)]$; $P_3 = [SS / H^+]$; and $P_4 = [(PAR * t^3) / (WD^3 * H^+)]$. F_{CH_4} , PAR , WD , ST , SS , C_p , H^+ , and t refer, respectively, to CH₄ emission flux, photosynthetic active radiation, water depth relative to soil surface, soil temperature, soil salinity, specific heat at constant pressure, hydrogen ion concentration, and time.

Principal component analysis

The first two PC's explained around 77% of the total data and system variance (Figure 8.2). Non-orthogonal (0 degree) orientations among P₁, P₂, and R₁ suggested high linkages. P₄ was moderately linked with R₁ and the linkage between the P₃, and R₁ were very weak. On the other hand, a moderate orientation between R₂ and (P₁, P₂, and P₃) implied the existence of moderate linkages between them.

Factor analysis

First two factors were extracted based on the initial eigenvalue criterion, eigenvalue >1 (Table 8.7). The high loadings of R₁, P₁, P₂, and P₄ on the first factor suggested their strong linkages. Moreover, R₂ loaded highly on 1st factor indicated its moderate linkage with P₁, P₂, and P₄. Although, P₃ loaded highly on factor 2, none of the response groups (R₁ and R₂) loaded significantly on this factor; suggesting P₃ was not an important predictor group.

Table 8.7: Latent factors extracted from the data matrix for F_{CH_4} . Bold indicates variables that loaded highly on each factor.

Dimensionless group	P ₁	P ₂	P ₃	P ₄	R ₁	R ₂
Factor 1	0.95	0.92	-0.03	0.76	0.83	-0.41
Factor 2	-0.04	-0.04	0.99	0.19	-0.14	0.17

Notes: $R_1 = [(F_{CH_4} * t) / (WD * SS)]$; $R_2 = [(F_{CH_4} * WD^2) / (PAR * t)]$; $P_1 = [(PAR * t^3) / (WD^3 * SS)]$; $P_2 = [(ST * C_p * t^2) / (WD^2)]$; $P_3 = [SS / H^+]$; and $P_4 = [(PAR * t^3) / (WD^3 * H^+)]$. F_{CH_4} , PAR, WD, ST, SS, C_p , H^+ , and t refer, respectively, to CH_4 emission flux, photosynthetic active radiation, water depth relative to soil surface, soil temperature, soil salinity, specific heat at constant pressure, hydrogen ion concentration, and time.

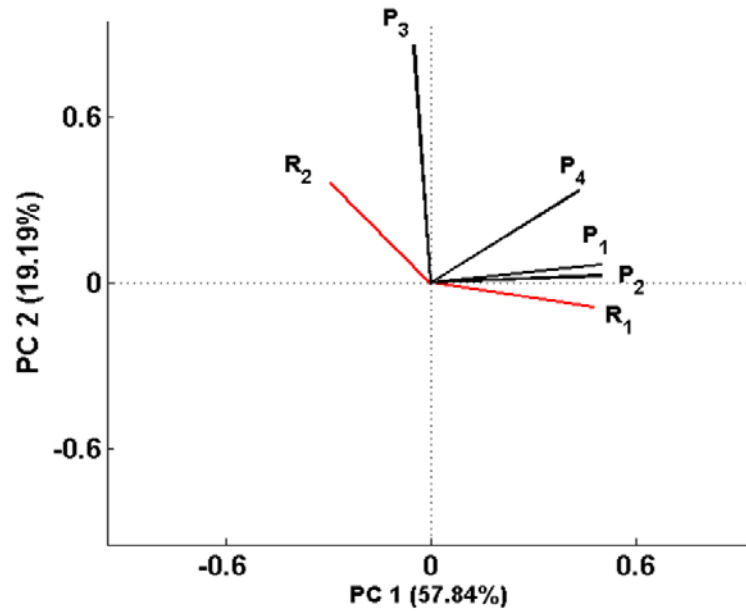


Figure 8.2: PCA biplots showing correlation strength and orientation of variables for wetland F_{CH_4} for all the study ponds of Waquoit Bay, MA. Percent variance explained by each PC is shown in parentheses. R_1 and R_2 are carbon flux group ($R_1 = [(F_{CH_4} * t) / (WD * SS)]$ and $R_2 = [(F_{CH_4} * WD^2) / (PAR * t)]$); P_1 and P_4 are the radiation-redox group ($P_1 = [(PAR * t^3) / (WD^3 * SS)]$ and $P_4 = [(PAR * t^3) / (WD^3 * H^+)]$); P_2 is the hydro-climatic group $[(ST * C_p * t^2) / (WD^2)]$; and P_3 is the salinity-redox group $[SS / H^+]$. F_{CH_4} , PAR, WD, ST, SS, C_p , H^+ , and t refer, respectively, to CH_4 emission flux, photosynthetic active radiation, water depth relative to soil surface, soil temperature, soil salinity, specific heat at constant pressure, hydrogen ion concentration, and time.

The results obtained from the data-analytics framework for CH_4 fluxes were not similar to the results of CO_2 fluxes. For CH_4 , both of the response variables (R_1 and R_2) were detected as the important response groups while formulating the scaling relations. As such both R_1 and R_2

can be expressed as a scaling function of P_1 and P_2 (Equation 2 and 3). However, identified important predictor groups of CH_4 were similar to the predictor groups that were identified for CO_2 flux.

$$\frac{F_{CH_4} * t}{SS * WD} = f \left[\left(\frac{PAR * t^3}{WD^3 * SS} \right) \left(\frac{ST * C_p * t^2}{WD^2} \right) \right] \quad (2)$$

$$\frac{F_{CH_4} * WD^2}{PAR * t} = f \left[\left(\frac{PAR * t^3}{WD^3 * SS} \right) \left(\frac{ST * C_p * t^2}{WD^2} \right) \right] \quad (3)$$

8.3.3 Examination of similitude and environmental regimes

CO₂ sequestration flux

Similitude, environmental regimes, and scaling relations across study ponds

The dimensionless predictor group, P_1 was plotted against the dimensionless response group, R_1 on a logarithmic scale (Figure 8.3a) for different study sites. To plot the data in the logarithmic scale, absolute values of the WD (when $WD < 0$) were considered. Two clear spatial regimes emerged from the plot that was not apparent in the dimensional domain. The lower regime ($P_1 < 1000000$) of the “radiation-redox, P_1 ” group had driven the CO_2 flux uptake of the sage lot salt marshes; this indicated sage lot salt marshes had sequestered CO_2 fluxes in a PAR limiting (relatively less available sunlight) and high anoxic- saline (high redox) condition. The Hamblin pond CO_2 sequestration followed a transitional regime between the high and low values of P_1 . For the other ponds (GP and EP) high regime ($P_1 > 1000000$) of “radiation-redox” (plenty of sunlight and less anoxic) was required for the CO_2 sequestration. Similarly, Sage lot salt marshes sequestered carbon for the low values of P_2 (hydro-climatic regime) (figure 9.3b); suggesting the Sage lot salt marshes can uptake CO_2 in a relatively low temperature and high flooding condition and other ponds needed relatively higher “hydro-climatic” regime for CO_2 uptake.

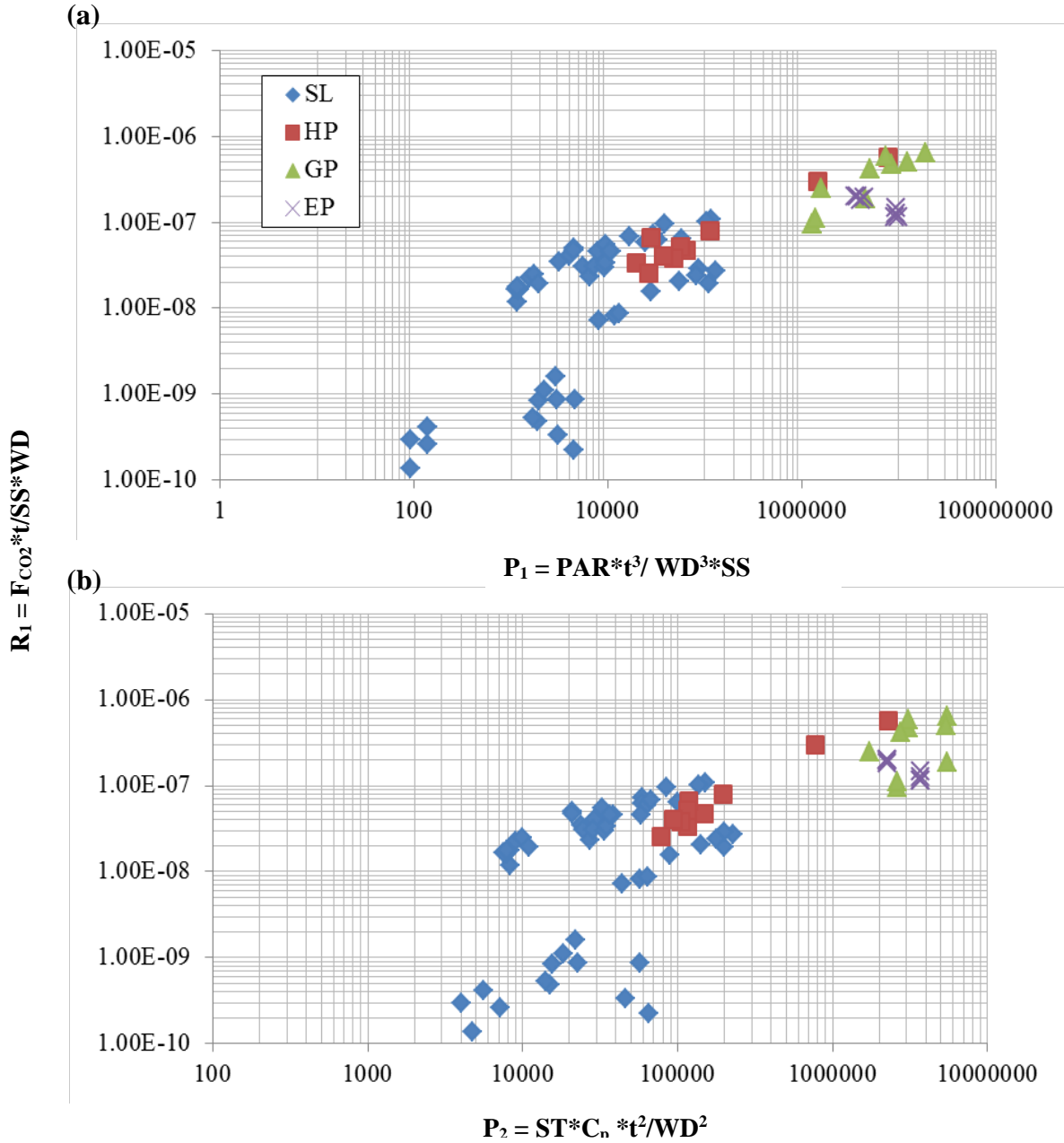


Figure 8.3: Plot of the predictor dimensionless groups ($\log_{10}P_1$ and $\log_{10}P_2$) with the response dimensionless group ($\log_{10}R_1$), revealing collapse of different variables on the power-law scaling relationships. R_1 is the CO₂ flux group $[(F_{CO_2} * t) / (WD * SS)]$; P_1 is the radiation-redox group $[(PAR * t^3) / (WD^3 * SS)]$; and P_2 is the hydro-climate group $[(ST * C_p * t^2) / (WD^2)]$.

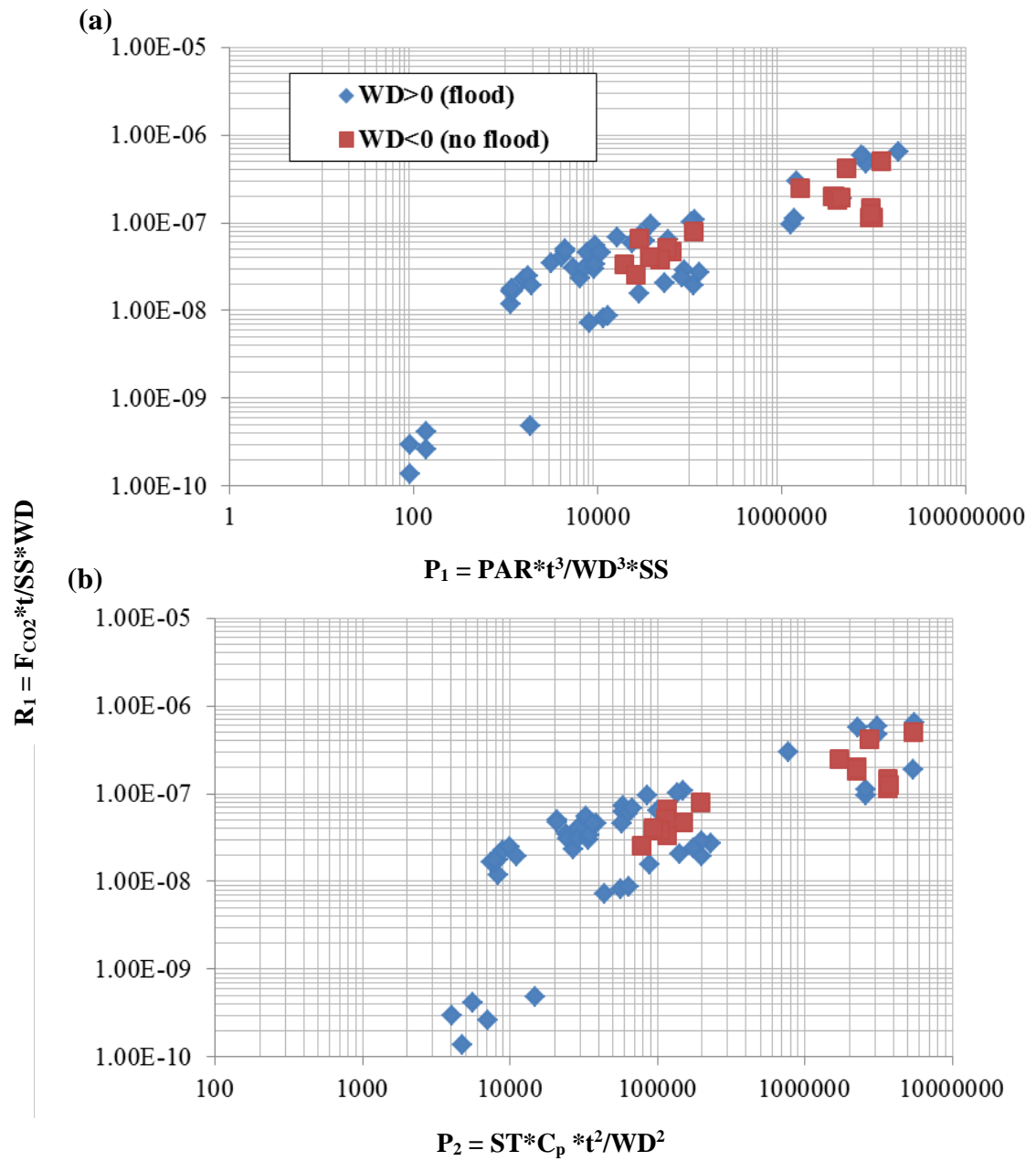


Figure 8.4: Plot of the predictor dimensionless groups ($\log_{10}P_1$ and $\log_{10}P_2$) with the response dimensionless group ($\log_{10}R_1$) for two different hydrological regimes (flood and no-flood), revealing collapse of different variables on the power-law scaling relationships. R_1 is the CO_2 flux group [$(F_{CO_2} * t) / (WD * SS)$]; P_1 is the radiation-redox group [$(PAR * t^3) / (WD^3 * SS)$]; and P_2 is the hydro-climate group [$(ST * C_p * t^2) / WD^2$].

Moreover, CO₂ sequestration rate was relatively less in lower P₁, and P₂ regime compare to the high P₁ and P₂ environment. The higher stress condition in the lower regime (PAR limiting, low temperature, high salinity, and flooding) is probably limiting the CO₂ sequestrating rates in the tidal wetlands.

Similitude, environmental regimes, and scaling relations for flood/non-flood condition

To investigate the potential existence of the hydrologic regime in the study sites, data were separated based on the flood (WD>0) and no-flood (WD<0) conditions. The P₁ and P₂ were then plotted separately against R₁ in a logarithmic scale (Figure 8.4a, b). Observations of both the hydrological conditions collapsed on top of each other, which indicated that functional relationship between the response and predictor numbers would be similar for both of the conditions. However, the lower and higher environmental regimes observed in Figure 8.3 was also a function of the hydrological regime since WD was included in both P₁ and P₂ numbers.

The shape factor values (b₁, b₂) were estimated from the data for both WD>0 and WD<0 cases by using the following power-law model to investigate the scaling relationships:

$$R_1 = P_1^{b_1} * P_2^{b_2}$$

The estimated parameter values were 3.46 and -4.71 for b₁ and b₂, respectively in the case of WD>0 and the model fitting was satisfactory (NSE: 0.58, RSR=0.64). However, for the WD<0 condition the b₁ and b₂ values were not similar (2.90 and -4.1 for b₁ and b₂, respectively), which suggested the presence of two different regimes govern by the hydrology. The poor fitting efficacy of the model in WD< 0 condition (NSE: -0.23, RSR = 1.1) also indicated that without a scaling factor the model prediction would be very uncertain. Moreover, relatively higher b₁ and b₂ values that were observed for WD>0 than the WD<0 conditions indicated a higher response of CO₂ sequestration with change in physical factors during completely submerged scenario.

CH₄ emission flux

Similitude, environmental regimes, and scaling relations across study ponds

The dimensionless predictor group, P_1 was plotted against both of the dimensionless response groups (R_1 and R_2) on a logarithmic scale (Figure 8.5a). Similar to the CO₂ observations, two clear regimes emerged in the plot for both R_1 and R_2 , which was not apparent in the dimensional domain. The lower regime of the “radiation-redox, P_1 ” group ($P_1 < 1000000$) showed strong functional relationship with R_1 and R_2 in the Sage lot site; this indicated Sage lot salt marshes followed a PAR limiting (relatively less available sunlight) and high anoxic- saline (high redox) condition. Hamblin pond CH₄ emission also followed the lower regime. The other ponds (GP and EP) were scaled in the higher regime of “radiation-redox” (High PAR and less anoxic) related to the CH₄ emission. Similarly, Sage lot and Hamblin pond salt marshes emitted CH₄ for the low values of P_2 (hydro-climatic regime) (Figure 8.5b) and other sites responded with the high values of P_2 . Comparing the collapse of the response groups, it was evident that both R_1 and R_2 can potentially be used for scaling relationships. It was also evident from the Figure 8.5a, b that R_1 increased with the increase of both P_1 and P_2 but R_2 decreased with the increase of P_1 and P_2 .

Similitude, environmental regimes, and scaling relations for flood/non-flood condition

To investigate the existence of the hydrologic regime in the study sites, data were separated based on the flood ($WD > 0$) and no flood ($WD < 0$) conditions. The P_1 and P_2 were then plotted separately against R_1 and R_2 in a logarithmic scale (Figure 8.6a, b). P_1 and P_2 did not show any scaling relationship with R_1 at $WD < 0$ condition based on the fitting efficiency ($b_1 = 3.86$, $b_2 = -5.66$, $NSE = -3.95$, $RSR = 2.15$). However, R_2 showed a significant evidence of scaling with P_1 and P_2 for $WD < 0$ condition ($b_1 = -1.83$, $b_2 = 1.35$, $NSE = 0.91$, $RSR = 0.29$). Completely opposite phenomenon was observed for $WD > 0$ condition. R_1 scaled very efficiently with P_1 and P_2 in the flooding condition ($b_1 = 5.07$, $b_2 = -7.05$, $NSE = 0.35$, $RSR = 0.80$) whereas R_2 did not scaled with P_1 and

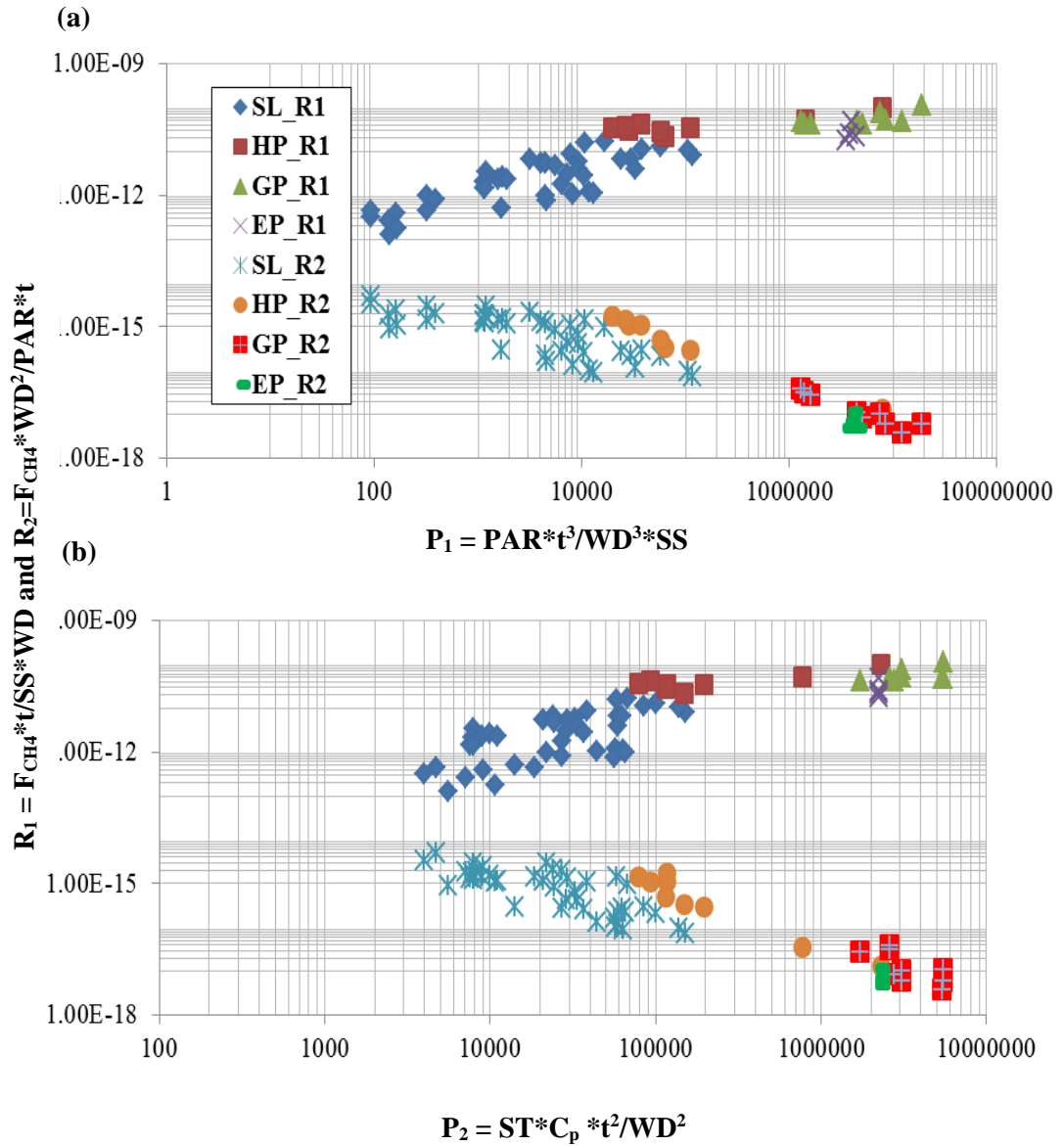


Figure 8.5: Plot of the predictor dimensionless groups ($\log_{10}P_1$ and $\log_{10}P_2$) with the response dimensionless groups ($\log_{10}R_1$ and $\log_{10}R_2$), revealing collapse of different variables on the power-law scaling relationships. R_1 and R_2 are the CH_4 flux group ($R_1 = [(F_{CH_4} \cdot t) / (WD \cdot SS)]$ and $R_2 = [(F_{CH_4} \cdot WD^2) / (PAR \cdot t)]$); P_1 is the radiation-redox group $[(PAR \cdot t^3) / (WD^3 \cdot SS)]$; and P_2 is the hydro-climate group $[(ST \cdot C_p \cdot t^2) / (WD^2)]$.

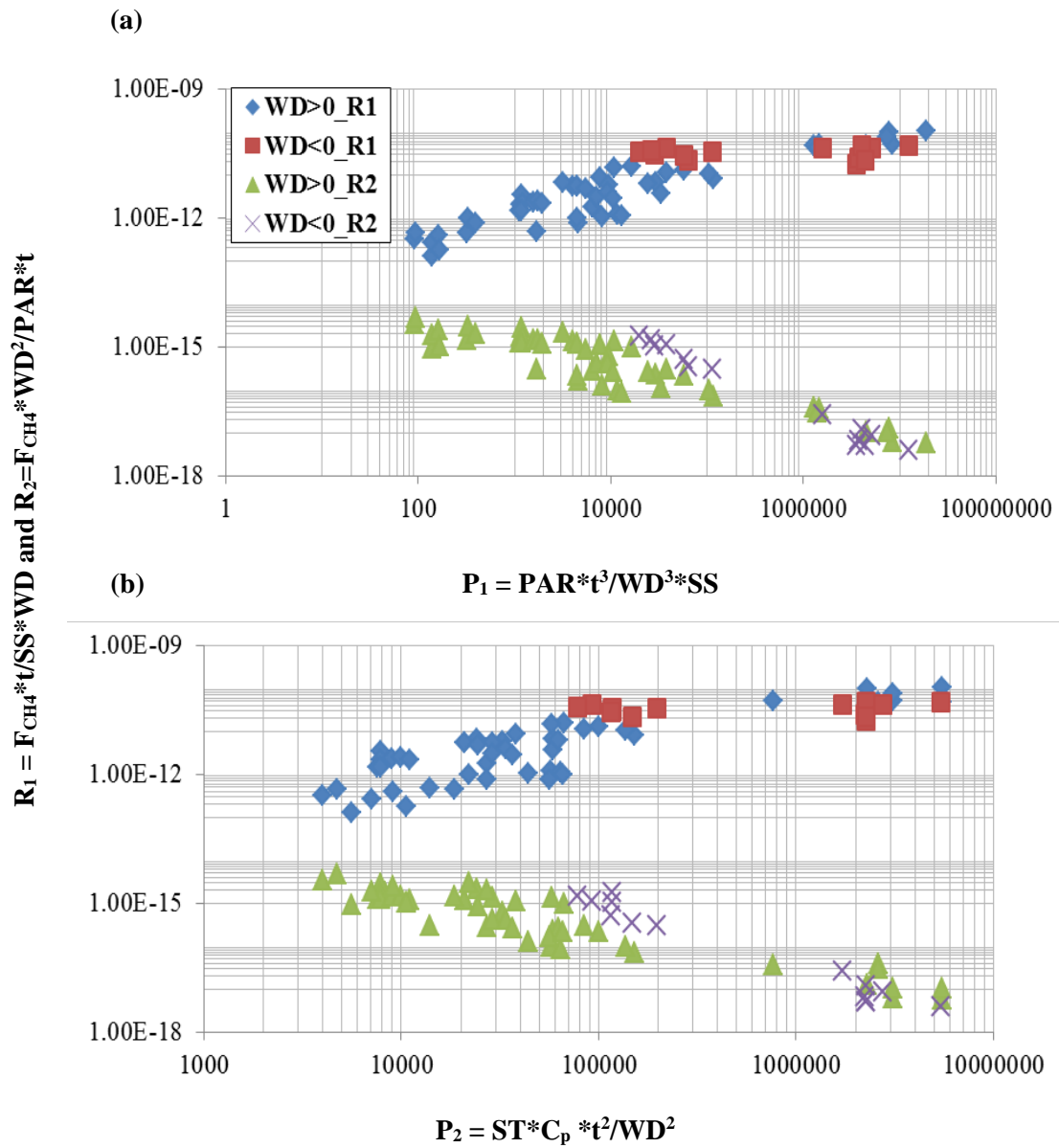


Figure 8.6: Plot of the predictor dimensionless groups ($\log_{10}P_1$ and $\log_{10}P_2$) with the response dimensionless groups ($\log_{10}R_1$ and $\log_{10}R_2$) for two different hydrological regimes (flood and no-flood), revealing collapse of different variables on the power-law scaling relationships. R_1 and R_2 are the CH_4 flux group ($R_1 = [(F_{CH_4} * t) / (WD * SS)]$ and $R_2 = [(F_{CH_4} * WD^2) / (PAR * t)]$); P_1 is the radiation-redox group $[(PAR * t^3) / (WD^3 * SS)]$; and P_2 is the hydro-climate group $[(ST * C_p * t^2) / WD^2]$.

P_2 as apparent from the model fitting measures ($b_1=0.81$, $b_2=-4.41$, $NSE=0.11$, $RSR=0.93$).

Overall, although different groups collapsed onto each other for both $WD>0$ and $WD<0$ condition, the scaling relationship between prediction and response variables was different for these conditions. These suggested different modeling scenarios should be explored by considering the hydrologic regime while predicting the CH_4 group.

8.3.4 Development of scaling laws and modelling

CO₂ sequestration flux

Since investigation on scaling and similitude for different spatial and hydro-climatic regimes indicated the existence of the two environmental regimes in salt marsh CO₂ uptake, two sets of models were developed based on the previously discussed cases of $WD>0$ and $WD<0$ to formulate scaling relationship and predict the response dimensionless group, R_1 from the important and meaningful predictor groups (P_1 and P_2) by using the following power-law equation:

$$R_1 = 10^{b_0} P_1^{b_1} * P_2^{b_2}$$

Case 1: modeling when $WD>0$ condition

Since the data were separated based on the position of the water table relative to the marsh surface, there is no need to mix the positive and negative values of the WD . The model (Equation 4) result was very impressive ($NSE=0.97$, $RSR=0.18$) (Table 8.8), and all the model parameters were statistically significant. The higher positive dominance of P_1 was observed whereas estimated parameter value of P_2 was negative. The scaling relation implied that dimensionless CO₂ flux group scaled positively with the radiation-redox group, P_1 and scaled negatively with the temperature-hydrology group, P_2 .

$$\frac{F_{CO_2} * t}{SS * WD} = 10^{-7.75} \left[\frac{PAR * t^3}{WD^3 * SS} \right]^{0.89} \left[\frac{ST * C_p * t^2}{WD^2} \right]^{-0.73} \quad (4)$$

Equation 4 can be written as a sole function of F_{CO_2} :

$$F_{CO_2} = 10^{-7.13} \left[\frac{SS * WD}{t} \right]^1 \left[\frac{PAR * t^3}{WD^3 * SS} \right]^{0.89} \left[\frac{ST * C_p * t^2}{WD^2} \right]^{-0.73} \quad (5)$$

The scaling relationship of the F_{CO_2} then can be expressed from Equation 5 as (considering t is fixed for a particular temporal scale)

$$F_{CO_2} \sim \frac{PAR^{0.89} SS^{0.11}}{ST^{0.73} C_p^{0.73} WD^{0.21}} \quad (6)$$

Equation 6 showed F_{CO_2} scaled to $PAR^{0.89}$, $SS^{0.11}$ and inversely scaled to $ST^{0.73}$ and $WD^{0.21}$.

Table 8.8: The estimated model parameters and model fitting efficacy for different modeling scenarios to predict CO_2 dimensionless number (shaded area is for $WD < 0$ condition).

Case	Parameter	Estimate	SE	t-stat	p-value	NSE	RSR
WD>0	b0	-7.75	0.20	-38.61	<0.001	0.97	0.18
	b1	0.89	0.06	15.23	<0.001		
	b2	-0.73	0.09	-8.45	<0.001		
WD<0	b0	-16.55	2.14	-7.72	<0.001	0.71	0.52
	b1	-1.65	0.46	-3.58	<0.01		
	b2	3.25	0.78	4.16	<0.001		

Note: $R_1 = 10^{b_0} P_1^{b_1} P_2^{b_2}$ is the predictive power law model for the prediction of F_{CO_2} . $R_1 = [(F_{CO_2} * t) / (WD * SS)]$; $P_1 = [(PAR * t^3) / (WD^3 * SS)]$; $P_2 = [(ST * C_p * t^2) / (WD^2)]$. F_{CO_2} , PAR , WD , ST , SS , C_p , and t refer, respectively, to net CO_2 sequestration flux, photosynthetic active radiation, water depth relative to soil surface, soil temperature, soil salinity, specific heat at constant pressure, and time. b_0 , b_1 and b_2 are the estimated model parameters.

Case 2: modeling when $WD < 0$ condition

The model (Equation 7) fitting was satisfactory (NSE=0.71, RSR=0.52) (Table 8.8), and all the model parameters were statistically significant (p-value<0.05). A comparison between b_1 and b_2 showed higher control of P_2 over P_1 on F_{CO_2} group contrast to the case 1 where P_1 had the higher

influence. Moreover, the dimensionless CO₂ flux group, R₁ scaled negatively with the radiation-redox group, P₁ and scaled positively with the temperature-hydrology group, P₂.

$$\frac{F_{CO_2} * t}{SS * WD} = 10^{-16.55} \left[\frac{PAR * t^3}{WD^3 * SS} \right]^{-1.65} \left[\frac{ST * C_p * t^2}{WD^2} \right]^{3.25} \quad (7)$$

Equation 7 can be written as a sole function of F_{CO₂}:

$$F_{CO_2} = 10^{-7.13} \left[\frac{SS * WD}{t} \right]^1 \left[\frac{PAR * t^3}{WD^3 * SS} \right]^{-1.65} \left[\frac{ST * C_p * t^2}{WD^2} \right]^{3.25} \quad (8)$$

The scaling relationship of the F_{CO₂} then can be expressed from Equation 8 as (considering t is fixed for a particular temporal scale)

$$F_{CO_2} \sim \frac{ST^{3.25} C_p^{3.25} SS^{2.65}}{PAR^{1.65} WD^{0.55}} \quad (9)$$

Equation 9 showed F_{CO₂} scaled to ST^{3.25}, SS^{2.65} and inversely scaled to PAR^{1.65} and WD^{0.55}.

CH₄ emission flux

Since the investigation on scaling and similitude on the range of spatial and hydro-climatic regimes indicated the existence of different regimes (based on high and low values of P₁ and, WD>0, and WD<0 conditions) in wetland CH₄ emission, we developed different sets of models by considering different regime (WD>0 and WD<0) scenarios to predict both of the response dimensionless groups (R₁ and R₂) from the important and meaningful predictor groups (P₁ and P₂) by using the following power-law equation:

$$R_1 = 10^{bo} P_1^{b1} * P_2^{b2}$$

$$R_2 = 10^{bo} P_1^{b1} * P_2^{b2}$$

This analysis is also helpful to identify appropriate response group from R_1 and R_2 that had profound scaling relationship with P_1 and P_2 under different regimes.

Case 1: modeling when $WD > 0$ (flood condition)

Modeling result with R_1 (Equation 10) was very good (NSE=0.93, RSR=0.26) (Table 8.9) and all the model parameters were statistically significant; whereas model developed using R_2 (Equation 11) showed moderate modeling accuracy (NSE=0.67, RSR=0.57). The model fitting for the $WD > 0$ condition suggested the scaling relations and predictive model could be competently developed by using R_1 . Although scale factor, b_0 of both the models were comparable, the shape factors were different for both R_1 and R_2 . The dimensionless CH_4 flux group, R_1 scaled positively with the radiation-redox group, P_1 and scaled negatively with the temperature-hydrology group, P_2 .

$$\frac{F_{CH_4} * t}{SS * WD} = 10^{-12.47} \left[\frac{PAR * t^3}{WD^3 * SS} \right]^{0.47} \left[\frac{ST * C_p * t^2}{WD^2} \right]^{-0.14} \quad (10)$$

$$\frac{F_{CH_4} * WD^2}{PAR * t} = 10^{-12.92} \left[\frac{PAR * t^3}{WD^3 * SS} \right]^{-0.21} \left[\frac{ST * C_p * t^2}{WD^2} \right]^{-0.32} \quad (11)$$

Equation 10 can be written as a sole function of F_{CH_4} :

$$F_{CH_4} = 10^{-12.47} \left[\frac{SS * WD}{t} \right]^1 \left[\frac{PAR * t^3}{WD^3 * SS} \right]^{0.47} \left[\frac{ST * C_p * t^2}{WD^2} \right]^{-0.14} \quad (12)$$

The scaling relationship of the F_{CH_4} then can be expressed from Equation 12 as (considering t is fixed for a particular temporal scale)

$$F_{CH_4} \sim \frac{SS^{0.53} PAR^{0.47}}{ST^{0.14} C_p^{0.14} WD^{0.13}} \quad (13)$$

Case 2: modeling when $WD < 0$ (no-flood condition)

With the $WD < 0$ condition, an opposite phenomenon emerged indicating the existence of two different hydrological regime in wetland CH_4 emission. R_2 was predicted (Equation 15) very successfully (NSE= 0.98, RSR=0.11) from P_1 and P_2 but the same predictor set did not able to predict the response group, R_1 (NSE=0.04, RSR=0.94) (Table 9) (Equation 14) (Table 8.9). All the estimated parameters for R_2 were statistically significant.

$$\frac{F_{CH_4} * t}{SS * WD} = 10^{-11.14} \left[\frac{PAR * t^3}{WD^3 * SS} \right]^{-0.16} \left[\frac{ST * C_p * t^2}{WD^2} \right]^{0.27} \quad (14)$$

$$\frac{F_{CH_4} * WD^2}{PAR * t} = 10^{-8.03} \left[\frac{PAR * t^3}{WD^3 * SS} \right]^{-1.12} \left[\frac{ST * C_p * t^2}{WD^2} \right]^{-0.38} \quad (15)$$

Equation 15 can be written as a sole function of F_{CH_4} :

$$F_{CH_4} = 10^{-8.03} \left[\frac{PAR * t}{WD^2} \right]^1 \left[\frac{PAR * t^3}{WD^3 * SS} \right]^{-1.12} \left[\frac{ST * C_p * t^2}{WD^2} \right]^{-0.38} \quad (16)$$

The scaling relationship of the F_{CH_4} then can be expressed from Equation 16 as (considering t is fixed for a particular temporal scale)

$$F_{CH_4} \sim \frac{WD^{2.12} SS^{1.12}}{PAR^{0.12} ST^{0.38} C_p^{0.38}} \quad (17)$$

8.4 Discussion

8.4.1 Biogeochemical-ecological similitudes and environmental regimes

Since many dimensionless groups can be formed from the different combinations of the variables, it was quite important to identify the most relevant and mechanistically explainable normalized groups that can explain a significant part of the normalized GHG flux groups. From the data-

Table 8.9: The estimated model parameters and model fitting efficiencies for different modeling scenarios to predict CH₄ dimensionless numbers.

Response group	Case	Parameter	Estimate	SE	t-stat	p-value	NSE	RSR
R ₁	WD>0	b0	-12.47	0.26	-47.23	<0.001	0.93	0.26
		b1	0.47	0.06	7.96	<0.001		
		b2	-0.14	0.09	-1.56	<0.05		
	WD<0	b0	-11.14	0.92	-12.10	<0.001	0.04	0.94
		b1	-0.16	0.28	-0.59	>0.05		
		b2	0.27	0.43	0.65	>0.05		
R ₂	WD>0	b0	-12.92	0.47	-27.26	<0.001	0.67	0.57
		b1	-0.21	0.07	-2.96	<0.01		
		b2	-0.32	0.15	-2.09	<0.05		
	WD<0	b0	-8.03	0.80	-10.08	<0.0001	0.98	0.11
		b1	-1.12	0.08	-13.18	<0.0001		
		b2	-0.38	0.14	-2.75	<0.05		

Notes: R₁ or R₂=10^{b0}*P₁^{b1}*P₂^{b2} is the predictive power law model for the prediction of F_{CH₄}. R₁=[(F_{CH₄}*t)/(WD*SS)]; R₂=[(F_{CH₄}*WD²)/(PAR*t)]; P₁=[(PAR*t³)/(WD³*SS)]; P₂ = [(ST*C_p*t²)/WD²]. F_{CH₄}, PAR, WD, ST, SS, C_p, and t refer, respectively, to CH₄ emission flux, photosynthetic active radiation, water depth relative to soil surface, soil temperature, soil salinity, specific heat at constant pressure, and time. b0, b1 and b2 are the estimated model parameters.

analytics framework, two most important normalized groups (radiation-redox group, P₁ and hydro-climate group, P₂) were identified to establish scaling relationship with the response groups of F_{CO₂} and F_{CH₄} (R₁ and R₂) (Equation 1, 2, and 3). For F_{CO₂}, both P₁ and P₂ demonstrated strong linkages with R₁ (Figure 8.1, Table 8.4, 8.5). However, P₁ and P₂ were strongly linked with R₁ and moderately linked with R₂ for F_{CH₄} (Figure 8.2, Table 8.6, 8.7), which suggested scaling relations should be investigated for both R₁ and R₂ while modeling F_{CH₄}. Equation 1, 2, 3 indicate normalized CO₂ and normalized CH₄ groups were the unknown functions (f₁, f₂, and f₃) of two dimensionless predictor groups of P₁ and P₂. The unknown functions were needed to be determined from the cause-effect interactions between the predictor groups and the response groups.

Plotting of R₁ against P₁ and P₂ for CO₂ unraveled an existence of the environmental and biogeochemical regimes that were previously hidden in the data (Figure 8.3a, b). Figure 8.3a, b

suggested that the dimensionless F_{CO_2} group, R_1 increases with the increase of P_1 and P_2 . However, it does not straightly imply that an increase of PAR and decrease of WD, and SS would increase the CO_2 uptake, although it is mechanistically understandable and well known that PAR facilitates CO_2 uptake. Rather a comparative ratio of radiation (PAR) and soil biogeochemical factors (WD and SS) determines how it interacts with the R_1 group. Similar concepts apply to the P_2 group. The fascinating mechanism that was unraveled from the plots was the identification of two different environmental and biogeochemical based regimes separated by the P_1 and P_2 threshold. Observations that cascade below the P_1 threshold of 1000000 were governed by the SL and HP sites whereas observations of GP and EP sites falls in the upper regime ($P_1 > 1000000$). Similarly, the P_2 threshold was 10000 that separated two regimes. The observations imply that CO_2 uptake of SL and HP site may be low because of a high biogeochemical stress and a low photosynthetic ability of the salt marshes in these sites. The soil temperature can also play an important role on this. On the contrary, the GP and EP ponds could be subjected to high photosynthesis and temperature, and low inundation and salinity.

However, to test the hypothesis of low photosynthesis, temperature, high salinity and inundation of SL and HP resulted lower CO_2 uptake in a lower environmental regime, the mean differences of PAR, ST, SS, and WD between the low environmental regime sites (SL, HP) and high regime sites (GP, EP) (Table 8.10) were compared. No significant mean differences were observed for PAR, ST, and SS. However, WD varies significantly between two regimes; for the low regime the mean WS was positive meaning the overall flooding condition, and for the high regime sites the mean water table was negative and very close to the marsh surface. This indicated inundation of the marsh plays a significant role in CO_2 uptake and the favorable situation for the CO_2 uptake occurs when the water table is close the marsh surface. To further test this hydrological dependency, the dimensionless groups (P_1 and P_2) were separated based on the hydrological regimes and then plotted against R_1 irrespective of sites (Figure 8.4a, b). Most of

the observations were in the lower regime for the completely submerged ($WD > 0$) condition, which supports our hypothesis of hydrological dependency. For the $WD < 0$ condition when uptake is relatively high, most of the observations fall in a transitional zone of low to high P_1 and P_2 , and in the high regime.

Analogous similitude based regimes were observed for the F_{CH_4} . Figure 8.5a, b revealed comparable threshold of P_1 and P_2 that define the environmental regimes. SL clearly falls in the lower zone and HP followed a transitional zone. GP and EP site were in the higher environmental zone. The F_{CH_4} group, R_1 was positively linked to the P_1 and P_2 . However, R_2 , which is a ratio of F_{CH_4} , WD , and PAR were negatively linked with P_1 and P_2 . To test the regime based separation of the sites, the mean of variables of the different regimes were compared (last two rows of Table 8.10). Similar to the observations on CO_2 uptake, CH_4 emissions was also found as a strong function of water depth. Water table very close to the surface or at the surface was the most favorable condition for the CH_4 emission rather than the completely submerged condition when CH_4 was scaled by SS and WD . However, when CH_4 was scaled by the PAR , emission was comparatively lower in the higher regime. The overall higher control of water depth and soil saturation on CH_4 emission fluxes was comprehensively documented in the literature (Cui et al., 2005; Cheng et al., 2007; Bubier and Moore 1994). The net CH_4 emission from the wetlands is the functions of CH_4 production (methanogenesis), CH_4 oxidation and CH_4 emission (methanotrophy) and all of these microbial-mediated processes are regulated by the water table dynamics as water table and soil saturation determine the reduced environment below the soil surface. An extremely reduced environment (negative redox potential), which is favorable for the CH_4 production and emission, occurs during the flood conditions. On the contrary, as the water subsides during ebb tide, CH_4 oxidization occurs in the anoxic zone of the soil surface (typically in the upper soil layer) that reduces the rate of CH_4 emission.

Table 8.10: Mean based comparison of the physical factors of CO₂ and CH₄ fluxes for low (SL and HP site) and high environmental regimes (GP and EP site) based on the observations.

GHG flux	Environmental Regime	PAR (μmol/m ² /s)	ST (oC)	SS (PPT)	WD (m)
F _{CO2}	Low (SL, HP)	1477.83	19.14	29.56	0.108
	High (GP, EP)	1432.391	22.82	31.23	-0.004
F _{CH4}	Low (SL, HP)	1381.96	20.06	28.96	0.119
	High (GP, EP)	1288.50	22.30	30.69	-0.002

8.4.2 Scaling laws and modeling

CO₂ sequestration flux

A robust dimensionless scaling relationship was established between the R₁ and the predictor groups (P₁ and P₂) for both WD>0 and WD<0 conditions. For the WD<0 condition, the model explained most of the variances of the R₁ (NSE=0.97). However, for WD<0 condition model explained 71% of the variances. Since WD<0 condition involves both sequestration and CO₂ emission from the CH₄ oxidation, the mechanism is much complex than WD>0, which explains the inability of the model in predicting the rest of the variances from only the physical variables.

The equation 4 shows the scaling relation in the dimensionless domain for WD>0 condition. The positive power of P₁ suggests a positive correlation with R₁ and the negative power of P₂ indicates an inverse scaling. However, the negative power of P₂ does not imply an increase of ST causes a decrease of R₂, since the regression model fitted dimensionless carbon flux number, R₁ with the ratio of different variables (dimensionless numbers) as a form of P₁ and P₂. However, an important scaling relation can be derived if both sides of the Equation 4 is multiplied by the scaling factor of R₁ that is (SS*WD/t). The multiplication and compilation of the shape factors of the parameters led to the Equation 6 that represents how each variable scales to the F_{CO2} for the WD>0 condition. The power law scaling of the variables indicated PAR and ST are the most dominating variables that scale with F_{CO2}. WD has the moderate scaling relation,

but SS did not show a significant scaling association with the F_{CO_2} . The high power of PAR (0.89) indicates that the CO_2 sequestration is mostly dominated by PAR. However, the negative value of ST resembles a stomatal closure at high temperature to reduce flux exchanges. Since the entire system is already anoxic and completely submerged, increase of WD may provide a signal to the plant to keep the stomata open.

Equation 7 expressed the scaling relationship of R_1 for the $WD < 0$ condition. Equation 9 showed how each of the independent variable scaled with the F_{CO_2} in the dimensional domain. The dominance of PAR decreases as the respiration part increases with the increase of aerobic layer in the upper part of the marsh surface. The increase of the power of SS in this condition is an indicator of the overall higher sensitivity of the gas exchanges with the salinity. Since salt marshes are tolerant to the salt stress, up to an optimum level of salinity, marshes maintain their normal productivity (Lewis and Weber 2002; Wang et al., 2006). The higher scaling dependence of $SS-F_{CO_2}$ may be described as a correlation perspective rather than a cause-effect viewpoint.

CH₄ emission flux

A robust dimensionless scaling relationship was established between the R_1 and the predictor groups (P_1 and P_2) for $WD > 0$ (Equation 10). The model explained 93% variances of the R_1 (NSE=0.93). However, for $WD < 0$ condition instead of R_1 , R_2 was better predicted by P_1 and P_2 (NSE=0.98) (Equation 14).

In $WD > 0$ condition, P_1 was positively linked with the R_1 whereas P_2 was negatively linked. The dimensionless scaling Equation 13) showed that F_{CH_4} scaled positively with SS and PAR and negatively with ST and WD; these observations contradict our mechanistic understanding on methane emission and indicate further investigation. For the $WD < 0$ condition, R_2 was negatively linked with both P_1 and P_2 Equation 17 showed F_{CH_4} scaled with WD positively (high WD leads to CH_4 production and emission). However, the negative scaling of

ST is hard to describe; the inclusion of Cp in the P₂ group to normalize ST could be responsible for the negative scaling of ST with F_{CH₄}. SS scaled negatively with F_{CH₄} indicates a potential reduction of CH₄ emission under high saline condition. Regarding the soil salinity, being a polyhaline tidal marsh the Waquoit Bay is subjected to the low CH₄ emission compare to the relatively less saline salt marsh systems. In the sulfate-rich salt marshes, methanogens have to compete with the sulfate-reducing bacteria for the production of CH₄ and that limit the CH₄ emission for the wetland to some extent (Poffenbarger et al., 2011); however, this relationship may vary from site to site depending on the adaptation mechanism of the salt marshes (Megonigal et al., 2004). The scaling relations obtained for F_{CH₄} with ST requires further investigation by using the measured flux data from different coastal wetlands to independently validate the finding.

8.5 Conclusions

The study provides a primary understating of the similitude, different regimes and scaling relationships of coastal wetland CO₂ sequestration and CH₄ emission fluxes. Two useful and important dimensionless GHG flux predictor groups that consist of physical drivers of the fluxes were identified from a diligent application of the Buckingham-Pie theorem. The existence of two different environmental regimes among the study wetland ponds were identified which are governed by the wetland hydrology and biogeochemistry. The study reveals that the coastal wetland system becomes most productive (high sequestration and less emission) when the water table position is at the surface. Different power-law based dimensionless scaling relations were obtained by fitting the observed data for flood and non-flood conditions for both the CO₂ sequestration and CH₄ emissions dimensionless numbers. The scaling relations were later expressed in a dimensional domain to find the GHG fluxes scaling relations with the physical drivers. The power-law based scaling of PAR was identified as the most dominating factor in

CO₂ sequestration whereas the position of the water table was the most influential scaling factor for the CH₄ emission.

The study revealed an evidence of biogeochemical similitude across different study ponds that are subjected to a moderate nitrogen gradient. Although two separate regimes were identified, the final scaling relationships were successfully determined by combining the data from both the regimes; supported our initial hypothesis of an existence of biogeochemical similitude across sites of similar ecosystem type. However, more investigation is needed by incorporating the data from different coastal salt marshes to independently calibrate and validate the observed scaling relations. A successful calibration and validation with different sites would leverage the potential spatial translation of the scaling relationships. On the temporal scale, the study is limited to the instantaneous fluxes that are in per second scale. However, since the time dimension is kept in the derived dimensionless groups, a temporal upscaling would not change the scaling relations, which is subjected to further research. Incorporation of the entire year data for consecutive years could validate the temporal robustness of derived scaling relations. This study could be considered as the 1st step in the domain of coastal wetland GHG flux exchanges that investigates the possibility of finding wetland biogeochemical similitudes and scaling relations from a vigorous data analysis and mining among different physical and biological factors.

References

- Abdul-Aziz, O.I., Wilson, B.N., Gulliver, J.S., 2007a. An extended stochastic harmonic analysis (ESHA) algorithm: Application for dissolved oxygen. *Water Resource Research* 43, doi:10.1029/2006WR005530.
- Abdul-Aziz, O.I., Wilson, B.N., Gulliver, J.S., 2007b. Calibration and validation of an empirical dissolved oxygen model. *Journal of Environmental Engineering* 133(7), 698-710.
- Bartlett, K., Bartlett, D., Harriss, R., Sebacher, D., 1987. Methane emissions along a salt marsh salinity gradient. *Biogeochemistry* 4 (3): 183-202.
- Bridgham, S. D., Megonigal, J. P., Keller, J. K., Bliss, N. B., Trettin, C., 2007. Wetlands. In: A.W. King, L. Dilling, G.P. Zimmerman, et al. (Eds.), *The First State of the Carbon Cycle Report (SOCCR): The North American Carbon Budget and Implications for the Global Carbon Cycle. A Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research.*
- Bubier, J. L., Moore, T. R., 1994. An ecological perspective on methane emissions from northern wetlands. *Trends in ecology & evolution*, 9(12), 460-464.
- Cheng, X., Peng, R., Chen, J., Luo, Y., Zhang, Q., An, S., ... Li, B., 2007. CH₄ and N₂O emissions from *Spartina alterniflora* and *Phragmites australis* in experimental mesocosms. *Chemosphere*, 68(3), 420-427.
- Chmura, G. L., Anisfeld, S. C., Cahoon, D. R., Lynch, J.C., 2003. Global carbon sequestration in tidal, saline wetland soils. *Global Biogeochemical Cycles* 17 (4): 1111.
- Conrad, R., 1989. Control of methane production in terrestrial ecosystems. In: M. O. Andrae and D. S. Schimel (Eds.), *Exchange of Trace Gases Between Terrestrial Ecosystems and the Atmosphere*, John Willy, NY, pp. 39-58.
- Cui, J., Li, C., Sun, G., Trettin, C., 2005. Linkage of MIKE SHE to Wetland-DNDC for carbon budgeting and anaerobic biogeochemistry simulation. *Biogeochemistry*, 72(2), 147-167.
- Duarte C. M., Middelburg, J., Caraco, N., 2005. Major role of marine vegetation on the oceanic carbon cycle. *Biogeosciences* 2: 1-8.
- Hargreaves, K. J., Fowler, D., 1998. Quantifying the effects of water table and soil temperature on the emission of methane from peat wetland at the field scale. *Atmospheric Environment* 32(19): 3275-3282.
- Howarth, R. W., 2014. A bridge to nowhere: methane emissions and the greenhouse gas footprint of natural gas. *Energy Science & Engineering*, 2(2), 47-60.
- Ishtiaq, K.S., Abdul-Aziz, O.I., 2015. Relative linkages of canopy-level CO₂ fluxes with the climatic and environmental variables for US deciduous forests. *Environmental Management*, 55 (4), 943-960.
- Kundu, P.K., Cohen, I.M., 2004. *Fluid Mechanics*. 3rd Edition, Elsevier Academic Press, NY.

- Lewis, M. A., Weber, D. E., 2002. Effects of substrate salinity on early seedling survival and growth of *Scirpus robustus* Pursh and *Spartina alterniflora* Loisel. *Ecotoxicology*, 11(1), 19-26.
- Megonigal, J. P., Mines, M. E., Visscher, P. T., 2004. Anaerobic metabolism: linkages to trace gases and aerobic processes. In: Schlesinger WH (ed) *Biogeochemistry*, Elsevier-Pergamon, Oxford, pp 317-424.
- Poffenbarger, H. J., Needelman, B. A., Megonigal, J. P., 2011. Salinity influence on methane emissions from tidal marshes. *Wetlands*, 31(5), 831-842.
- Sebacher, D. I., Harriss, R. C. , Bartlett, K. B., Sebacher, S. M., Grice, S. M., 1986. Atmospheric methane sources: Alaskan tundra bogs, an alpine fen, and a subarctic boreal marsh. *Tellus* 38B(1):1-10.
- Smith, K. A., Ball, T., Conen, F., Dobbie, K. E., Massheder, J., Rey, A., 2003. Exchange of greenhouse gases between soil and atmosphere: interactions of soil physical factors and biological processes. *European Journal of Soil Science* 54(4): 779-791.
- Stocker, T. F., Qin, D., Plattner, G. K., Tignor, M., Allen, S. K., Boschung, J., ... Midgley, B. M., 2013. IPCC, 2013: climate change 2013: the physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change.
- Walter, B. P., Heimann, M., 2000. A process-based, climate-sensitive model to derive methane emissions from natural wetlands: Application to five wetland sites, sensitivity to model parameters, and climate. *Global Biogeochemical Cycles* 14: 745– 765.
- Wang, Q., Wang, C. H., Zhao, B., Ma, Z. J., Luo, Y. Q., Chen, J. K., Li, B., 2006. Effects of growing conditions on the growth of and interactions between salt marsh plants: implications for invisibility of habitats. *Biological Invasions*, 8(7), 1547-1560.
- Wang, Z., Zeng, D., Patrick, W.H., 1996. Methane emissions from natural wetlands. *Environmental Monitoring and Assessment* 42(1): 143-161.
- Warnaars, T. A., Hondzo, M., Power, M. E., 2007. Abiotic controls on periphyton accrual and metabolism in streams: Scaling by dimensionless numbers. *Water Resources Research* 43,W08425, doi:10.1029/2006WR005002.
- Waquoit bay national estuarine research reserve website, 2015
<http://www.waquoitbayreserve.org/>
- West, G. B., Brown, J. H., Enquist, B. J., 2001. A general model for ontogenetic growth. *Nature* 413: 628-631.
- Yu, Z., 2006. Power laws governing hydrology and carbon dynamics in northern peatlands. *Global and Planetary Change*, 53(3), 169-175.

Chapter 9: Summary and conclusions

The dissertation investigated the organizing principles of ecosystem GHG fluxes, and developed empirical and data-based mechanistic models for obtaining robust predictions of GHG fluxes from diverse ecosystems. The spatial scale of the models varies from site-specific to ecosystem to regional scales, encompassing different forest and wetland ecosystems. The research provided crucial understanding and insights into the linkage and emergence patterns of ecosystem GHG fluxes. The predictive modeling tools for CO₂ and CH₄ fluxes are expected to assist ecologists, engineers and stakeholders with a proper management of ecosystem carbon and associated services under a changing climate and environment.

Chapter 2 and 3 examined the first hypothesis and introduced a scaling-based modeling framework in an entirely empirical domain to predict the diurnal cycles of terrestrial net ecosystem exchange (NEE) from a single reference observation. The scaled-based, empirical modeling framework was developed and successfully applied to predict the diurnal cycles of terrestrial NEE in **chapter 2**. The developed model incorporated a parsimonious set of five parameters, which exhibited spatiotemporal robustness by collapsing into narrow ranges with no apparent trends among the different days and sites. The model parameters were also insensitive to the individual, crucial process drivers (e.g., photosynthetically active radiation, temperature, and soil water content) of NEE. Model calibrations illustrated good fitting efficiency (coefficient of determination, $R^2 = 0.68$ to 0.88) and accuracy (ratio of root-mean-square error to the standard deviation of observations, $RSR = 0.35$ to 0.57) for different forests. Upscaling of the individual forest-specific models to that at the generalized deciduous forest level was performed by averaging the temporal ensemble of model parameters among all study forests. The applicability of the model as a gap-filling method was also tested by comparing the accuracy and efficiency of the proposed model with the different available NEE gap-filling methods.

The scaling-based modeling framework was also applied to robustly predict the diurnal cycles of wetland NEE from the corresponding single reference observations in **chapter 3**. The scaled, dimensionless model was evaluated with different five year data for a coastal freshwater wetland (Florida Everglades Taylor slough marsh; 2008-12) and an inland shrub wetland (Lost Creek, Wisconsin; 2004-08), as available in the AmeriFLUX network. The model calibrations and validations showed good efficiency ($R^2 = 0.64$ to 0.77 ; Nash-Sutcliffe Efficiency, $NSE = 0.63$ to 0.77) and accuracy ($RSR = 0.48$ to 0.61) in predicting the hourly NEE for the study wetlands. Therefore, chapter 2, 3, and 4 successfully tested the first hypothesis by developing predictive models of diurnal cycle of terrestrial and wetland NEE; the models were robust in space and time, and presented important examples of scaling application in ecological engineering.

Development of a parsimonious data-driven model of GHG fluxes demands a selection of the most relevant and meaningful drivers. A systematic, multivariate data-analytics approach was proposed in this research to determine the relative linkages and controls of different environmental and climatic drivers with the GHG fluxes. The framework was examined to determine the physical linkages of instantaneous CO_2 fluxes of deciduous forests in **chapter 4**. The study showed that the ‘radiation-energy’ component had the strongest linkages with the canopy-level CO_2 fluxes, whereas the ‘temperature-hydrology’ component had low to moderate carbon flux linkages. The study did not find any significant CO_2 linkages with the ‘aerodynamic’ component. On average, the ‘radiation-energy’ component showed around 5 and 8 times stronger carbon flux linkages than that of the ‘temperature-hydrology’ and ‘aerodynamic’ components, respectively. Moreover, all the study sites exhibited comparable linkages and groupings with the CO_2 fluxes, indicating a potential for developing a generalized, data-driven predictive model of CO_2 fluxes for the deciduous forests. Moreover, the explanatory empirical models, as developed to quantify the linkages, showed relatively good model accuracy and efficiency ($R^2 = 0.55$ to

0.81; RSR = 0.44 to 0.67), indicating the usefulness of the multivariate analytics models for estimating and predicting GHG fluxes.

In **chapter 5**, the applicability of the data-analytics methodology for diverse terrestrial ecosystems was tested. Six different AmeriFLUX ecosystems were used as the study sites to determine the relative linkages of the climatic and environmental variables with the NEE. The study found that the ‘radiation-energy’ group demonstrated the strongest linkages (60% contribution) with the NEE; exhibiting on average 4 (17% contribution), 5 (14% contribution) and 8 (9% contribution) times stronger linkages than that of ‘temperature-hydrology’, ‘ambient carbon concentration’ and ‘aerodynamic’ components, respectively. The developed explanatory models showed satisfactory accuracy (training RSR= 0.48-0.66, validation RSR= 0.50 to 0.65) and efficiency (training $R^2= 0.57-0.77$, validation $R^2= 0.58-0.75$) to link NEE with its physical drivers. The findings from this chapter demonstrated the ability and scope of data-analytics to select the most meaningful and major drivers of GHG fluxes in developing a parsimonious predictive model for diverse ecosystems, including wetlands.

Utilizing the data-analytics framework, predictive models of CO_2 and CH_4 fluxes from coastal wetlands were developed in **chapter 6 and 7**, respectively. Dominant controls of radiation (photosynthetically active radiation) and temperature, as well as a moderate control of soil salinity on daytime CO_2 exchanges were observed for the salt marshes. In contrast, soil temperature, sediment salinity, and water depth were identified as the most important physical drivers of wetland CH_4 emission. Predictive, data-based mechanistic models of instantaneous CO_2 and CH_4 fluxes were then developed as nonlinear (power-law) functions of the most important variables influencing the gas fluxes. The models exhibited very good fitting efficiency ($R^2=0.79-0.89$) and accuracy (RSR=0.32-0.45) in predicting the wetland CO_2 and CH_4 fluxes. The models were presented in an Excel spreadsheet as a handy ecological engineering tool for

estimations and predictions of GHG fluxes and carbon storage in tidal marshes of the New England region. Estimations of the net ecosystem carbon balance (NECB) using the tool will guide preservation and restoration project activities in relation to 'blue carbon'. Overall, chapter 5-8 tested the second hypothesis by pursuing objectives 2 and 3.

Chapter 8 tested the third hypothesis and provided an understanding of the biogeochemical-ecological similitudes, critical environmental regimes, and scaling laws for wetland CO₂ sequestration and CH₄ emission fluxes. The study identified two important dimensionless physico-chemical groups to predict the dimensionless GHG flux groups. Two distinct environmental regimes were found based on wetland hydrology and biogeochemistry. The study found dominant control of water table in regulating GHG fluxes; the New England tidal marshes appeared most productive (high sequestration and low emission) when the water table position is at the ground surface. The dimensionless scaling relations of CO₂ and CH₄ fluxes with the climatic and biogeochemical drivers were obtained through a power-law based regression analysis. The study revealed an evidence of biogeochemical-ecological similitude for wetland GHG fluxes across the Waquoit Bay study ponds, which represented a low to moderate nitrogen and a high salinity and hydrologic (water table) gradients. This study could be considered as the first step in the research domain that investigates the possibility of finding wetland biogeochemical-ecological similitudes and scaling laws for GHG fluxes by involving data-analytics of relevant physical and biological factors.

Future research is necessary to expand the spatial and temporal boundary of the developed models by incorporating data from different sites along a gradient of nitrogen, climate, canopy height, vegetation, inundation, and salinity. Furthermore, model generalization should be achieved through a successful validation. Investigation of the temporal robustness of the developed data-analytics framework is another potential avenue of further research. Since the

dissertation tests the usefulness of the data-analytics framework up to hourly scales, future works could incorporate a multi-time scale feature (daily, weekly, monthly, and annually) to investigate the multi-scale linkages of climatic and environmental variables with the GHG fluxes. Usefulness of the developed models in estimating the annual balance (i.e., temporal upscaling) should also be tested in future research. The topic of discovering similitude, scaling laws and organizing principles for ecosystem GHG fluxes clearly merits further investigations and many future efforts.

The data-analytics based study is an emerging field in the domain of ecological sciences and engineering. This research presented a novel and systematic approach for data-based mechanistic knowledge-building, model generalization, and parsimonious predictive modeling of terrestrial and wetland GHG fluxes. Utilization of scaling and similitudes within a data-analytics framework enriches the existing mechanistic knowledge to develop scale-invariant models for obtaining robust predictions of ecosystem GHG fluxes under a changing climate and environment. The predictive models developed in this dissertation will act as ecological engineering tools to aid the development of appropriate GHG offset protocols for setting design and monitoring guidelines for ecosystem restoration and maintenance projects.

VITA

KHANDKER SAQIB ISHTIAQ

EDUCATION AND WORKING EXPERIENCE

- 2004-2009 B.Sc., Water Resources Engineering
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
- 2009-2011 M.Sc., Water Resources Engineering
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
- Junior Engineer, River Engineering Division
Institute of Water Modeling
Dhaka, Bangladesh
- 2012-2014 GPSC Award for Conference Presentation,
Florida International University, Miami
- 2011 -2015 Doctoral Candidate, Civil and Environmental Engineering
Florida International University
Miami, Florida, Unites States
- Teaching and Research Assistant
Department of Civil and Environmental Engineering
Florida International University, Miami, Florida
- 2015 Dissertation Year Fellowship (DYF)
University Graduate School
Florida International University
- Provost Award (GSAW Scholarly Forum Winner),
Florida International University, Miami

PUBLICATIONS AND PRESENTATIONS

Abdul-Aziz, O. I., Ishtiaq, K. S., 2014. Robust Empirical Modeling of Dissolved Oxygen in Small Rivers and Streams: Scaling by a Single Reference Observation, *Journal of Hydrology*, 511, 548-557, doi: <http://dx.doi.org/10.1016/j.jhydrol.2014.02.022>.

Ishtiaq, K.S., Abdul-Aziz, O.I., 2015. Relative linkages of canopy-level CO₂ fluxes with the climatic and environmental variables for US deciduous forests, *Environmental Management*, 55 (4), 943-960, DOI: 10.1007/s002670140437

Ishtiaq, K.S., Abdul-Aziz, O.I., 2015. Scaling-based robust empirical modeling of terrestrial net ecosystem exchange: Application for deciduous forests, Agriculture and Forest Meteorology, under revisions.

Abdul-Aziz, O. I., Ishtiaq, K. S., 2015. Comparative linking of climatic and ecohydrological variables with the net ecosystem exchange (NEE) of diverse ecosystems using a data-analytics framework, under internal review.

Ishtiaq, K.S., Abdul-Aziz, O. I., 2012. Application of Scaling for Robust Modeling of Dissolved Oxygen in Streams and Rivers , AGU Fall Meeting, San Francisco, CA.

Abdul-Aziz, O. I., Ishtiaq, K.S., 2012. Scaling for Robust Assessment and Predictions of Aquatic Ecosystem Health, AGU Fall Meeting, San Francisco, CA.

Ishtiaq, K.S., Abdul-Aziz, O. I., 2013. Quantitative linking of dominant environmental drivers and fluxes with vertical CO₂ fluxes of eight deciduous forests, AGU Fall Meeting, San Francisco, CA.

Abdul-Aziz, O. I., Ishtiaq, K.S., 2013. Relative linkages of environmental drivers and fluxes with net ecosystem exchanges of six diverse terrestrial ecosystems, American Geophysical Union Fall Meeting, San Francisco, CA.

Abdul-Aziz, O. I., Wilson, B.N., Ishtiaq, K.S., 2013. Application of scaling in ecological engineering, 98th ESA Annual Meeting, Minnesota.

Ishtiaq, K.S., Abdul-Aziz, O.I., 2014. Scaling for Robust Empirical Modeling and Predictions of Net Ecosystem Exchange (NEE) from Diverse Wetland Ecosystems, AGU Fall Meeting, San Francisco, CA.

Abdul-Aziz, O.I. and Ishtiaq, K.S., 2014. Robust Modeling and Predictions of Greenhouse Gas (GHG) Fluxes from Coastal Wetland Ecosystems, AGU Fall Meeting, San Francisco, CA.

Abdul-Aziz, O.I., Ishtiaq, K.S., 2014. Empirical modeling and predictions of greenhouse gas fluxes from coastal wetlands: A data-analytics approach, Joint Aquatic Sciences Meeting 2014, Portland, Oregon (Invited).

Abdul-Aziz, O. I., Ishtiaq, K.S., 2015. Developing User-friendly Models of Greenhouse Gas (GHG) Fluxes from Coastal Wetlands, SWS Annual Meeting, Rhode Island.

Egan, K., Moseman-Valrierra, S., Abdul-Aziz, O.I., Ishtiaq, K.S., 2015. Greenhouse gas emissions across salt marsh zones in Waquoit Bay, Massachusetts, USA: testing relationship to plant and environmental properties, New England Estuarine Research Society Spring 2015 Meeting, Bristol, Rhode Island.