© 2022 Leila Constanza Hernandez Rodriguez

INFORMATION DRIVEN CAUSAL STRUCTURE, EVOLUTION, AND FUNCTIONALITY OF COMPLEX BIOSPHERE - ATMOSPHERE SYSTEMS

BY

LEILA CONSTANZA HERNANDEZ RODRIGUEZ

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Civil Engineering
in the Graduate College of the
University of Illinois Urbana-Champaign, 2022

Urbana, Illinois

Doctoral Committee:

Professor Praveen Kumar, Chair and Director of Research
Professor Francina Dominguez
Professor Carl Bernacchi
Professor Atul Jain
Professor Hoshin Gupta, University of Arizona

Abstract

At the biosphere-atmosphere interface, an eddy covariance flux tower acts as an integrator of information from several processes occurring in the ecosystem at different space and time scales. These high frequency observations, typically at 10Hz or 20Hz, contain information about multiple variables such as wind speed and direction, humidity, temperature, water vapor (H_2O), and carbon dioxide (CO_2). These variables together capture a complex system of interactions where time-dependent non-linear interdependencies sustain the dynamics and are able to influence the behavior of each component. Such interdependencies arise due to the dynamic connectivity of the system through which fluctuations in one variable propagate to other variables at a lagged time. At the tower, the near-instantaneous measurements are averaged for the estimation of the exchange of fluxes using the eddy covariance technique, resulting in continuous measurements of the ecosystem-scale scalar exchange of heat, water, and CO_2. These observations are used in research and management worldwide, mainly to understand the temporal and spatial variations at the ecosystem scale. However, these observations also carry information about the influence of the composition inside its flux footprint, the temporally dynamic source/sink area that contributes to the measured fluxes. It also contains causal information regarding the dynamics of multivariate interactions at turbulence scales, which has not yet been explored in a multivariate context.

In this dissertation, the primary goal of the research is to understand how the information carried in the observations from an eddy covariance tower captures changes in the structure, evolution, and functionality of the system at different scales. The specific goals of this dissertation are to (i) determine in what way the composition of the heterogeneous organization of the land cover inside the flux footprint contributes to the observations measured at the tower; (ii) develop a framework for the study of the evolution of the causal structure of multivariate interactions at the turbulence scale by exploiting the high frequency measurements, and (iii) determine the relationship between the evolution of the causal structure and its relationship to the system's functionality.

Based on data from a 25m tall eddy covariance flux tower that sees maize (corn) and soybean fields in the agricultural Midwest of the US, which comprises a "patchwork quilt" of different crops, we determine how this mosaic crop pattern impacts the exchange of water, heat, and CO_2 fluxes measured at the tower. Then, we determine the relative contribution

of the different land cover types to the total flux. Our hypothesis is that the human-induced spatial reorganization of the landscape results in an exchange of heat, water and CO_2 that is different from the exchange that would result if the vegetation species were randomly (statistically homogeneously) distributed. To address this hypothesis, we estimate the spatial extent contributing to every flux measurement from April 2016 to April 2019 and combine them with an ecohydrological model to obtain the temporally varying contribution of fluxes for different land covers. Then we estimate how they contribute to the overall flux, which depends on how the crop fields are distributed. We find that the tower mostly "sees" fields located between 168m and 268m away although areas as far away as several kilometers contribute to the observations. As a result of crop rotation, maize fields contributed more than soybean fields during the 2016 and 2018 growing seasons, and vice versa during 2017. We compare our results to a hypothetical case where all vegetation is randomly distributed on the surface. We find that the mosaic organization of crop fields that the tower "sees" account for up to 24.5% more CO_2 flux (F_c) than if the vegetation was randomly distributed on the landscape. We find that the knowledge of footprint contributions combined with ecohydrological model results helps explain how the contributions from different crops to the observed fluxes at the tower vary from year to year.

To address the second goal, we explore how the eddy covariance measurements of high frequency (i.e., 10Hz) multivariate data also contain information about the interdependency between interacting variables such as horizontal and vertical wind speed, humidity, temperature, and CO_2. These measurements offer an opportunity to move beyond the traditional spectral analyses to explore causal dependency among them at turbulence scales. We visualize the temporal dependencies among lagged variables that illustrate its dynamics in a Directed Acyclic Graph (DAG) representation for time series. Causal associations among variables are represented by links in a network of interactions denominated "causal structure", whose representation is a DAG. We estimate the network of inter-dependencies using DAGs from multivariate high frequency time series and quantify its evolutionary dynamics. Our hypothesis is that there are different types of causal structures embedded in the day-time turbulence of the land-atmosphere exchange that can be recognized as the system evolves. To test our hypothesis, we develop a novel approach to find patterns of similar behavior using the DAGs, which is based on a distance-based classification to characterize the structural differences between the DAGs and a k-means clustering approach that selects an unbiased number of clusters. We explore a large range of feasible dynamics of the system by using data during a clear sky day and during the solar eclipse in 2017. We compare the sequences of DAGs to investigate structural differences in causal interactions among the variables. We find that in this multivariate system, temperature, horizontal and vertical wind velocities, along with scalars such as CO_2 and H_2O, create a causal structure of interdependencies that sustain information flow and evolves inside a range of dynamical behavior instead of being in a fixed configuration. Our results show well-defined clusters of similar structural connectivity and the emergence of patterns of similar dynamics while the system evolves. This causal history approach for investigating the evolution of multivariate dynamics, combined with the novel approach of the evolution of the causal structure provides a methodological framework to understand how dependence in turbulence is manifest at high frequencies.

Finally, we explore the relationship between the causal structure and its associated function, using the representative DAGs in each cluster of similar dynamics. We determine if functional differences are a reflection of structural differences, or in spite of structural differences, there are no functional differences. We based our analysis on the hypothesis that causal relationships are affected through information flow, and the interaction arising due to information flow sustains the dynamics of the ecohydrological system at high frequency, its functionality. To test our hypothesis, we build upon our causal structure analysis to estimate the information flow in each cluster using a multivariate causal history approach. We characterize functionality as the nature of interactions as discerned through redundant, unique, and synergistic components of information flow. Redundant information is overlapping information provided by multiple sources to a target, unique information is only provided by a single target, and synergistic information is provided only when two or more sources are present together. We find that in turbulence at the biosphere-atmosphere interface, the variables that control the dynamic character of the atmosphere as well as the thermodynamics, are driven by non-local conditions, while the scalar transport associated with CO_2 and H_2O are mainly driven by short-term local conditions.

In the absence of a causal dynamical law that is able to describe the behavior of complex systems in the geosciences, we develop a novel approach to account for the evolution of the structural causal model. The approach implemented here for investigating the functionality of the system through the evolution of information flow, supported by the novel approach of the evolution of the system's causal structure, provides a methodological framework to understand how a complex system evolves. In our implementation we use the information carried in the continuous time series of observations at a flux tower to better understand the behavior of multivariate turbulence at the biosphere-atmosphere interface. With the increasing availability of observational data at different temporal and spatial scales, and with the rapid advance of data-driven approaches based in observations, the implementation of causality in machine learning and other artificial intelligence (AI) applications would help us recognize causal patterns of behavior and improve the understanding of the dynamics of complex systems to inform Earth systems models, prediction, and ultimately decision-making.

To my family

Acknowledgments

I would like to thank my advisor, Professor Praveen Kumar, for his insights and for trusting me in this pursuit, and for his continuous encouragement and support over the past years. I am grateful for his “I won’t stop anyone’s curiosity!" which encouraged me to take as many opportunities as I possibly could to learn and to grow as a researcher and as a person during my Ph.D. journey. I also would like to thank my committee members: Prof. Francina Dominguez, Prof. Carl Bernacchi, and Prof. Atul Jain, professors at the University of Illinois Urbana - Champaign, and Prof. Hoshin Gupta, Regents Professor at the University of Arizona, for their advice and contributions towards this work. Likewise, thanks to my fellow colleagues in the Information Theory in the Geosciences reading group, which we initiated in Fall 2020, for all enlightening discussions.

I am grateful to our Kumar-nation research group: Allison Goodwell for her guidance on research. Peishi Jiang for helping me navigate the information theory waters. For all our memorable conversations and friendship, Qina Yan, Susana Roque-Malo, Meredith Richardson, Esther Lee, Kunxuan Wang, Frank Holcomb, Afeefa Rahman, Tarun Agrawal, Aramide Moronfoye, and Harsh Sarju Shah. I am deeply thankful to Susana for bringing Rocky into my life. I am also grateful to our “hydro-family" at the Ven Te Chow Hydrosystems Laboratory, especially to Cindy, Tony, and Julian Chen, Maria Kontari, Dimitrios Fytanidiz, and Kevin Wallington, for all fun moments and friendship.

My deepest gratitude goes to my beloved family: my husband Andrés Felipe, my mom Maria Antonia, my dad Luis Enrique, my sister Angela Sofia, and my grandma Maria Antonia. Thank you for encouraging me to dream big. I have also been lucky enough to have a wonderful dog by my side while writing this dissertation and for most of my life. To Locky, Sandy, and Frida, who were all joy and from heaven accompany me... and to Río, for being an amazing little one who came home right before the pandemic started... I suspect Río does not know that I can actually live far away from a laptop... Dear puppy, I will make it up to you big!

Finally, thank you to everyone who makes research in science possible. I gratefully acknowledge the support from NSF Grant EAR-1331906 for the Critical Zone Observatory for Intensively Managed Landscapes *(IMLCZO)*, a multi-institutional collaborative effort, NSF grant OAC-1835834, NSF Grant EAR-2012850 for the Critical Interface Network for Intensively Managed Landscapes *(CINet)*, DOE ARPA-E Smartfarm DE-AR0001225, and the Grainger

College of Engineering at the University of Illinois at Urbana-Champaign. I especially thank Steve Sargent at the Illinois State Geological Survey for taking care of the instruments at the flux tower at Goose Creek and for manually retrieving the high frequency data from the data logger every month. To all of them, thank you for caring and for helping to keep our world rolling.

Table of contents

List of Abbreviations . . . . . **x**

List of Symbols . . . . . **xii**

Chapter 1 Introduction **1**
1.1 Background 2
1.2 A unique network of observational sites for the study of critical interfaces 9
1.3 Research questions and contribution 10
1.4 Organization 11

Chapter 2 Inside the flux footprint: Understanding the role of organized land cover heterogeneity on land-atmospheric fluxes **13**
2.1 Introduction 13
2.2 Study Site 16
2.3 Methods 17
2.4 Results and discussion 24
2.5 Summary and conclusions 35

Chapter 3 Causal interaction in high frequency turbulence at the biosphere-atmosphere interface: Structural behavior **39**
3.1 Introduction 39
3.2 Evolution of causal structure 42
3.3 Results 50
3.4 Discussion 60
3.5 Conclusions 62

Chapter 4 Causal interaction in high frequency turbulence at the biosphere-atmosphere interface: Structure - function coupling **64**
4.1 Introduction 64
4.2 Characterizing causal dependency 66
4.3 Results 72
4.4 Discussion: Duality between causal structure and functionality 83
4.5 Conclusions 83

Chapter 5 Conclusions **85**
5.1 Role of organized land cover heterogeneity inside the flux footprint 85
5.2 Evolution of the causal structure of interactions in high frequency turbulence 86
5.3 Evolution of causal structure - function coupling 88
5.4 Avenues for future research 89

References **91**

Appendix A Supplementary information **102**
A.1 Using a one-dimensional flux footprint approach when estimating the role of the organized land cover heterogeneity 102

A.2 Hydroclimatic variables measured at Goose Creek flux tower site 112
A.3 MLCan parameters 113
A.4 LE & F_c coupled behavior 114
A.5 A toy example of using the $K\left(G_l, G_m\right)$ metric 115
A.6 Testing the performance of the $K\left(G_l, G_m\right)$ metric 116
A.7 Estimating G_{full} without contemporaneous links 117
A.8 Cross-covariance of the time series in the DAGs of the cluster's centroids 118

List of Abbreviations

CO_2	Carbon dioxide molar concentration $[kg/m^3]$
H_2O	Water vapor molar concentration $[g/m^3]$
T	Air temperature $[°C]$
WS	Horizontal wind speed $[m]$
U_z	Vertical wind speed $[m]$
LE	Latent heat flux $[W/m^2]$
H	Sensible heat flux $[W/m^2]$
F_c	Carbon dioxide flux $[\mu mol/m^2 s]$
CDF	Cumulative Density Function
CMI	Conditional Mutual Information
DAG	Directed Acyclic Graph
DI	Dunn Index
EC	Eddy covariance
FFT	Fast Fourier transform
IT	Information theory
$\mathscr{k}nn$	$\mathscr{k}$-nearest-neighbor
MI	Mutual Information
MIT	Momentary Information Transfer
MITP	Momentary Information Transfer along causal Path
MIWTR	Momentary Information Weighted Transitive Reduction
MPID	Momentary Partial Information Decomposition
PID	Partial Information Decomposition
PDF	Probability Density Function
TE	Transfer Entropy

TR	Transitive Reduction
USDA	United States Department of Agriculture
WTR	Weighted Transitive Reduction

List of Symbols

X_t	Subprocess X at time t
$\vec{X}_t$	Multivariate process $\vec{X}$ at time t
$\vec{X}_t^-$	Causal history, which refers to all the historical states of $\vec{X}$ prior to time t
$H(X_t)$	Shannon's entropy. Uncertainty of a variable X_t
$H(X_t\|Y_t)$	Conditional entropy. The uncertainty of X_t that remains given the knowledge of Y_t
$I(X_t;Y_t)$	Mutual information between X_t and Y_t
$I(X_t;Y_t \mid Z_t)$	Conditional mutual information between X_t and Y_t given Z_t
τ	Positive time lag
$Y_{t-\tau} \to Z_t$	Directed edge that links two nodes $Y_{t-\tau}$ and Z_t , where τ is a positive time lag
$C_{X_{t-\tau}\to Z_t}$	Causal path from a source node $X_{t-\tau}$ to the target node Z_t in a DAG for time series
$C_{\vec{V}\to Z_t}$	Union of multiple causal paths from source nodes $\vec{V}$ to Z_t in a DAG for time-series
P_{X_t}	Parents of node X_t in a DAG for time series
$P_{C_{\vec{V}\to Z_t}}$	Parents of $C_{\vec{V}\to Z_t}$ in a DAG for time series
$I^{TE,\,\text{full}}_{Y\to Z}(\tau)$	Transfer entropy from the previous τ states of Y to Z_t
$I^{TE}_{Y\to Z}(\tau)$	Transfer entropy from a specific state of $Y, Y_{t-\tau},$ to Z_t

$I^{MIT}_{X_{t-\tau}\to Z_t}$	MIT from node $X_{t-\tau}$ to Z_t through a directed edge
$I^{MITP}_{X_{t-\tau}\to Z_t}$	MITP along causal path from node $X_{t-\tau}$ to Z_t through $C_{X_{t-\tau}\to Z_t}$
$\mathscr{T}$	Information flow from the whole causal history
$\mathscr{J}$	Information flow from the immediate causal history
$\mathscr{D}$	Information flow from the distant causal history
$\vec{W}$	Condition set for MPID
$\vec{W}_\tau$	Condition set for estimating $\mathscr{J}$ with immediate and distant histories separated by τ
R	Redundant information in PID
S	Synergistic information in PID
U	Unique information in PID
R_c	Redundant information in MPID
S_c	Synergistic information in MPID
U_c	Unique information in MPID
$R_{\mathscr{T}}$	Redundant information from self- and cross-dependencies of the causal history
$S_{\mathscr{T}}$	Synergistic information from self- and cross-dependencies of the causal history
$U_{self,\mathscr{T}}$	Unique information of the self-dependence in the causal history
$U_{cross,\mathscr{T}}$	Unique information of the cross-dependence in the causal history
$R_{\mathscr{J}}$	Redundant information from self- and cross-dependencies of the immediate causal history
$S_{\mathscr{J}}$	Synergistic information from self- and cross-dependencies of the immediate causal history
$U_{self,\mathscr{J}}$	Unique information of the self-dependence in the immediate causal history
$U_{cross,\mathscr{J}}$	Unique information of the cross-dependence in the immediate causal history

$R_{\mathscr{D}}$	Redundant information from self- and cross-dependencies of the distant causal history
$S_{\mathscr{D}}$	Synergistic information from self- and cross-dependencies of the distant causal history
$U_{self,\mathscr{D}}$	Unique information of the self-dependence in the distant causal history
$U_{cross,\mathscr{D}}$	Unique information of the cross-dependence in the distant causal history
$\mathscr{I}$	Interaction information
$\mathscr{I}_{\mathscr{D}}$	Interaction information of the distant causal history
$\mathscr{I}_{\mathscr{J}}$	Interaction information of the immediate causal history
$G = (V, E)$	Directed Acyclic Graph
E	Collection of edges in a DAG
V	Collection of vertices in a DAG
i, j	Two vertices of V where $i < j$ if $i \in V$ precedes $j \in V$
$e = (i, j)$	An edge e connecting two vertices i and j , while $f = (j, i)$ is e with reversed direction
$G = \left(\vec{X}_{t+1}^{-}, E\right)$	DAG representation for time series, where each node refers to the state at time t of a variable in $\vec{X}_t$ and E is the collection of edges
$K^{(p,q)}(G_l, G_m)$	Distance metric used to compare a pair of DAGs, G_l and G_m. Alternatively shown as $K(G_l, G_m)$ after the definition of the parameters p and q
G_{empty}	DAG with no edges
G_{full}	Fully connected DAG without contemporaneous edges
k	Number of clusters resulting from the k-means approach
C	Cluster of DAGs
$\delta(C_i, C_j)$	Inter-cluster distance between clusters C_i and C_j
Δ	Intra-cluster distance

D_r Degree of relative change

D_c Degree of connectivity

Chapter 1

Introduction

At the biosphere-atmosphere interface, the ecosystem-scale exchanges of heat, water, carbon dioxide, CO_2, and other gases are measured using the eddy covariance technique [1]–[3] in hundreds of flux tower stations around the world (e.g., AmeriFlux and FLUXNET) [4], [5]. The observations provided from flux towers support research and management applications worldwide, primarily to understand the temporal and spatial variations that are associated with the fluxes from plot to ecosystem scale [4]. Given the worldwide concern of the rapid change of the earth's atmosphere and biosphere since the industrial revolution, the study of the behavior of the CO_2 concentration, along with the other greenhouse gasses, continues to be an important topic of multi-disciplinary research. At the tower, near-instantaneous measurements (i.e., typically at 10Hz or 20Hz) are used for the estimation of the exchange of fluxes, which results in continuous time series of variables of high frequency data of wind velocity, temperature, and scalar concentrations of water vapor (H_2O) and CO_2. Using the eddy covariance technique [6], estimates of latent heat (LE), sensible heat (H), and CO_2 fluxes (F_c) at a temporal resolution from 15-minutes to 60-minutes are obtained through averaging of the high frequency measurements. Besides the magnitude and variability of the measured variable, these observations also carry information about the influence of the composition inside the flux footprint of the tower, the temporally dynamic source/sink area that contributes to the measured fluxes. They also contain information about the dynamics of the complex system shaped by causal interactions at turbulence scales, which has not yet been explored in a multivariate context. Therefore, these observations provide us an opportunity to move beyond traditional approaches of spectral and correlation analyses.

The information about the system that is carried in the variables at high frequency and in the averaged flux data, are useful to infer different elemental aspects of the system in which the tower is immersed. For example, CO_2 observations are the result of the ambient concentration in the atmosphere modulated by local uptake (e.g. photosynthesis) and release (e.g. autotrophic and heterotrophic respiration) in areas inside the tower's flux footprint, the dynamically varying

spatial distribution of source/sink areas contributing to the observed flux. To explore how these variation contribute to the observed values of CO_2, here we use the information about the ecosystem that is carried by the observations. However, CO_2 is only one of many variables in a multivariate ecohydrological complex system and its behavior is influenced by its multiple interaction with other components of the system. In nature, multivariate interactions among components sustain the dynamics of complex systems. For example, the complexity in the brain can be seen as giving rise to the emergence of consciousness due to the multitude of interactions [7]. In the stream chemistry system, the time series of multivariate data reflect $1/f$ fractal signatures of long-term dependencies due to the complex interactions occurring inside of the catchment [8], [9]. In this dissertation, we use the information encoded within the observations that captures the complex interactions of the biosphere-atmosphere exchange at different scales. First, we refer to the information that is carried by the 15-minutes observations regarding the land cover heterogeneity. Later, we use the high frequency data to understand the evolution of the causal structure of the interdependencies among variables in turbulence, represented as directed acyclic graphs (DAGs) [10], and their coupling with the functionality of the system.

In this dissertation, the primary goal is to understand how the information carried in the observations from an eddy covariance tower captures changes in the structure, evolution, and functionality of the system at different scales. The specific goals of this dissertation are to (i) determine in what way the composition of the heterogeneous organization of the land cover inside the flux footprint contributes to the observations measured at the tower; (ii) develop a framework for the study of the evolution of the causal structure of multivariate interactions at the turbulence scale by exploiting the high frequency measurements, and (iii) determine the relationship between the evolution of the causal structure and its relationship to the system's functionality.

1.1 Background

1.1.1 Towards flux footprint representativeness

The flux observations obtained using the eddy covariance technique [6], typically of latent heat (LE), sensible heat, (H), and CO_2 fluxes (F_c), are used primarily to understand their temporal and spatial variations [4]. These observations support research and management applications worldwide [4], [5]. However, the dynamic source/sink areas from where the fluxes originate are variable and usually not part of the data, which contrast with the fixed and explicit spatial extent of Earth system models and remote-sensing products that rely on flux tower observations for validation and parameterization [11], [12]. Therefore, the flux footprint, the temporally dynamic source/sink area that contributes to the measured fluxes, provides relevant and necessary information to contextualize the tower observations, and to determine how accurately the observations reflect upon the ecosystem where the tower is located [11], [13], [14]. This is a well-known aspect of using eddy covariance data, but it is frequently overlooked in terms of its implications for the

interpretation of the observations, which incentivizes our study.

Starting with the guidelines proposed by Pasquill [15] to identify the source areas, much more research has contributed to the development of the currently available techniques to define the flux footprint. The initially limited use of flux towers in quasi-homogeneous terrains in the 1980s, started gaining popularity to explore other not so idealized terrains supported by the rapid development of computers for processing the data and for performing flux footprint analysis [14]. Vesala, Kljun, Rannik, *et al.* [16] and Leclerc and Foken [14], provide an extensive review of approaches in terms of footprint modeling, including: analytical model, Lagrangian stochastic particle dispersion model, large-eddy simulations, and ensemble-averaged closure models. In the case of a one-dimensional footprint estimation [17], the associated fetch will shift direction and area at every time step, as the fetch points towards the prevalent upwind direction. In this context, the fetch corresponds to the distance from the tower when describing the footprint [6]. In Appendix A.1, we implemented the one-dimensional estimation of the flux footprint proposed by Schuepp, Leclerc, MacPherson, *et al.* [17] to determine the role of the land-cover heterogeneity on the flux observations. The beauty of the simplicity of this one-dimensional flux footprint approach allowed a first and clear connection between the surface roughness estimation and the resulting fetch. However, it assumes neutral atmospheric conditions and its implementation was recommended for homogeneous terrains. However, only a few eddy covariance sites are located in a homogeneous landscape. The two-dimensional flux footprint approach proposed by Kljun, Calanca, Rotach, *et al.* [18], which allows the estimation of the flux footprint under a wide variety of stability conditions in the atmosphere and changes in the surface roughness, offers a more appropriate method to estimate the distribution of the concentration or flux density [11], [13], [14]. This method is used in our study presented in Chapter 2.

In this dissertation we argue that beyond the magnitude and temporal variability of the measured flux, the observed data carries the influence of the spatial heterogeneity within the flux footprint. Therefore, we look for the source/sink areas from the underlying ecosystem that influence on the flux observations of the tower and focus on the physical characteristics of the land cover composition that is encoded in the observations. Looking for attributes inside the dynamic flux footprint, we find changes in the composition of the landscape that the tower can see which impact the response of the flux exchange, and consequently the observations. For a tall tower, the flux footprint is on the order of several square kilometers, which varies as a function of the measurement height, wind direction, and atmospheric stability [18]. A tall tower would be able to see the heterogeneous composition of the landscape from an ecosystem scale across different directions, which has important implications on the observations. An important aspect when analysing ecohydrological systems refers to the role of vegetation on flux exchange, which is a key component that acts as a regulator of various biochemical cycles. At this interface, vegetation connects several above and below ground processes, which impact the observations at the tower. While vegetation takes up and releases CO_2 during photosynthesis and respiration, respectively, different types of vegetation respond differently to prevailing hydrobiogeochemical conditions, which play a significant role on the overall system response. Therefore, the CO_2 measured at the tower must be

modulated by the composition of the landscape and their dynamics inside the flux footprint.

Previous studies have applied temporally varying footprint weights to construct statistical models for land cover fluxes from the spatially aggregated flux tower data [13], [19]–[21]. Some of those efforts rely on other sources of in situ data [22], [23]. However, a flux tower that is designed to provide observations of a single crop is unable to provide information about the role of heterogeneous organization of crop fields in observed fluxes. Therefore, when studying heterogeneous landscapes, many previous studies have relied on data from nearby towers for which flux footprints cover a specific vegetation type [22]–[24] or retrieved data from flux chamber measurements [13]. However, in highly heterogeneous systems with mixed vegetation and soil wetness, these sparse measurements might not be representative of the ecosystem behavior but reflect the particular conditions of a single source/sink plot, which potentially creates a bias in the analysis in the case of highly heterogeneous landscapes. Therefore, when hundreds of plots are located inside several square kilometers size dynamic flux footprints, these previous techniques may not reveal the average behavior of each land cover type inside the tower flux footprint for long periods. Another approach relies on the use of remote sensing or aircraft data to estimate fluxes from certain plots on the flight transect [25]. However, the temporal resolution of the flux data provided by these sources usually does not allow for long periods of analysis. Therefore, we require an alternative framework to inform the flux footprint analysis in terms of the surface source/sink flux at the ecosystem scale and for multiyear periods to allow for inter-annual comparative analyses.

In the case of a tall tower that sees an agricultural landscape in the Midwestern US [26], [27], maize (corn) and soybean fields create a "patchwork quilt" of land cover types. During the growing season, a patchy mosaic of different crops is the dominant feature, which partially hides other sources of heterogeneity such as soil properties and micro-topographic variability [25], [28]. In Chapter 2, we focus on the role of the "organized land cover heterogeneity", which we define as the human-induced spatial organization of the landscape, on the fluxes exchanged between the landscape and the atmosphere when measured at a 25m tall eddy covariance tower. Essentially, the observed fluxes at any time interval (e.g., 15min) arise from a mixture of contributions from different crop types in the dynamically changing flux footprint. To disentangle the contribution of each crop type, we need additional information. This is provided by an ecohydrological model. Knowing the composition of the heterogeneous land cover inside the flux footprint at each time interval, we use this additional information to determine the relative contribution of different land cover types to the total flux measured by the tower.

1.1.2 Process networks and equivalent DAGs

The early exploration of the effects of the solar radiation on atmospheric gases at the Earth surface by Foote [29] offers a first interpretation of the relationship between water vapor and CO_2 on air temperature [30]. From Foote experiments, we not only know that the air temperature is higher in damp air than in dry air, but those first observations

allowed the recognition of dependencies, such as that a greater amount of CO_2 in the atmosphere increases Earth's temperature. The early understanding of the effects of heat absorption by atmospheric gases has been relevant over the last century [31], leading to the exploration of the effects of the accelerated increase of greenhouse gases on the Earth and its implications on ecosystems from climatic, biological, economical, and social perspectives [32]–[35]. From an ecohydrological perspective, previous studies have investigated the effects of the magnitude of the greenhouse gases and/or air temperature increase on a variety of natural and agricultural systems, to better understand and even anticipate their effect on ecological services [36]–[38]. However, the understanding of the causal interdependence among variables in complex ecohydrological systems in which CO_2, water vapor, and other gases are active components that interact with other hydroclimatological variables, continues to be a topic of interest.

A particular case of complex system's self-organization [39], the emergent behavior that results from nonlinear interactions of components of a system, can be seen in the interaction of ecohydrological variables, such as air temperature, wind speed, and carbon dioxide concentration [40]. In general, the study of perturbations that propagate through the interactions across components of a system, such as water, ecosystems, climate, and human systems, among others, lead the emergence of patterns of similar behavior [41]. The variability in the connectivity among components of a system can be seen as information flow between the different variables along with the associated time lag [40], in which higher order information theory metrics provide deeper insights. An advancement towards the understanding of such interactions using hydroclimatological data for applications in ecohydrology was to measure the strength and time lag of the directional information flow between two variables using Transfer Entropy (TE) [42], [43], to establish time dependent dependency and feedback relationships. TE refers to the Conditional Mutual Information (CMI) between a source node and the target that is conditioned on the history of the target [44], [45]. The estimation of asymmetric information flow among all possible pairs of variables using TE creates a process network [44], [46] (Figure 1.1a), which considers information flow as a reduction in predictive uncertainty of one variable due to the knowledge of another variable that is lagged in time [40]. In the process network each time series variable is a node, and nodes are linked through time dependencies for the analysis of process interaction at short time lags [40]. In process networks, nodes exhibit deterministic or stochastic behavior due to functional dependencies or random influences, respectively [45]. However, TE does not account for intermediate or multiple drivers that may account for the information flow, a property that is not desirable given the nature of the multivariate system where many interactions can occur to determine the state of a target. A further step is reached by estimating the contributions from multiple source nodes to a target node using metrics based on 3d probability distribution functions [47], where the shared information between a pair of nodes can be decomposed into synergistic, unique, and redundant components (SUR, respectively) [48]. In a multivariate case a process network can be partitioned into the SUR components [47]. The unique information is the information that a single source node shares with the target node. Redundant information is shared between any source node and the target node. Finally, the synergistic information is obtained only when any two sources are known together [47]. The

SUR framework allows the construction of ecohydrologic process networks where multiple nodes are considered as targets and sources of synergistic, unique, and redundant information at different time scales [47], [49]. A further step refers to the estimation of the whole system behavior, also using information-theoretic metrics for characterizing the information flow to a target from two sources as well as of the entire causal history [50], which uses a Directed Acyclic Graph (DAG) representation for time series to represent the dynamics of the system (Figure 1.1b).

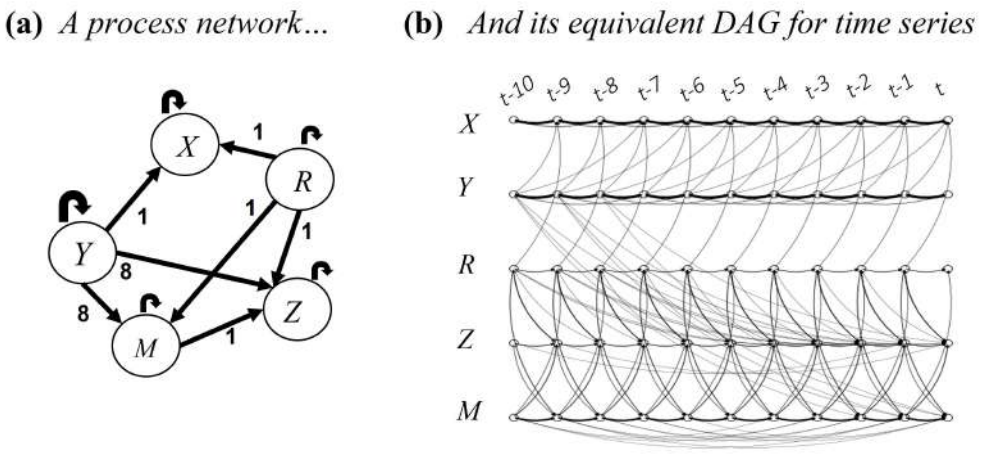

Figure 1.1: Graphical representation of the dynamics of a multivariate complex system in the form of (a) process network consisting of nodes (comprise of variables) that influence other nodes at time lags; and (b) directed acyclic graph (DAG) for time series that captures these temporal dependencies. In the process network the number next to the edge shows the temporal lag associated with the causal influence between any pair of linked variables. In the DAG for time series we visualize all dependencies of the system over time.

1.1.3 Causal interactions, inference, and evolution of DAGs

We use Directed Acyclic Graphs (DAGs) for time series [43], [51]–[53] to represent the information flow from one variable to another, as carried by the propagation of fluctuations, which is an indicator of causal dependency [54], [55]. A DAG for time series is a conceptual representation of a network structure that consists of a set of vertices and edges, alternatively called nodes and links, respectively, where an edge is directed from one vertex to another in a set of ordered pairs (Figure 1.1b). A DAG is “acyclic" in that the links never form a closed-loop [10]. In this dissertation we follow the notation in Jiang [56] to represent a multivariate complex system as a DAG for time series. Each of the variables in a multivariate ecohydrological system can be influenced by previous states of the variable itself or by other variables in the system at a lagged time [56]. For a complex system with N variables, $\vec{X}_t = \{X_t, Y_t, Z_t, \ldots\}_N$, which varies in time t, the influence from the system to the current state of any variable $Z_t \in \vec{X}_t$ accounts for all interactions in the previous dynamics of the system (Figure 1.1b). All the prior dynamics constitute the causal history, $\vec{X}_t^- = \left\{\vec{X}_{t-1}, \vec{X}_{t-2}, \vec{X}_{t-3}, \ldots\right\}$. In a DAG each node refers to a temporal variable $Z_t \in \vec{X}_{t+1}^-$. A directed edge linking two nodes $Y_{t-\tau}$ and Z_t, $Y_{t-\tau} \rightarrow Z_t$,

refers to a direct causal influence from $Y_{t-\tau}$ to Z_t, where τ is a positive time lag. We refer the reader to Chapters 3 and 4, where we provide a detailed description of the information-theoretical metrics for the estimation of a DAG for time series. In this dissertation, a DAG for time series represents a structural causal model [10], which we refer to as "causal structure".

To estimate the causal structure, we use the notion of "causal inference", which refers to the temporal asymmetric relationship among variables in the multivariable set [10]. Causal inference analyzes the response of an affected variable when the cause of the affected variable is changed [57]. One characteristic of this causal relationship is that the effect cannot precede the cause [10], [56]. Backward causation, the idea that there may be cases where the effect temporally, but not causally, precedes its cause [58], is not possible in our representation of the dynamics of the system when using DAGs. Here, the cause and the effect are placed in time so that the cause precedes its effect temporally [58]. Another feature is the asymmetric influence between the variables, meaning that the source-target dependency among variables is generally not the same when the order is reversed. We argue that causal interdependency among variables can be characterized as informational and is distinct from correlative relationships. Correlation refers to a temporal symmetric or synchronization in the behavior between variables, while associations arise due to the effect of a known or unknown common driver [10]. Causal influence captures the influence in one variable resulting from a perturbation that propagates from another variable. For instance, CO_2 and water vapor may show a strong correlative behavior because H_2O is spelled from plant stomates at the same time that CO_2 is absorbed during photosynthesis. However, humidity in the atmosphere may exert a causal influence by determining the amount of moisture that can diffuse from the stomates to the atmosphere. When we use information-theoretic metrics to determine the reduction of uncertainty in a target variable, given the knowledge of one variable or all the variables at different states in its past, which we discuss in Chapters 3 and 4, then we are estimating the causal interdependency of their dynamics and finding the propagation of fluctuations captured using informational measures or variations among variables.

Given that many systems evolve through periods of slow change of its dynamics sporadically supported by bursts of activity [59], in this dissertation we propose a novel approach for studying the evolution of the causal structure when using DAGs for time series, which is presented in Chapter 3. Given that the properties of a complex system change as a result of the intrinsic system dynamics [59], we focus on the quantitatively characterization of the evolution of its causal structure by determining changes in the connectivity of consecutive DAGs. In Chapter 4 we determine the relationship between the causal structure and the system's functionality. We define functionality as the nature of interactions as discerned through the SUR components using the causal history approach [9], [50].

1.1.4 Towards causality in multivariate turbulence

Once described as “the most important unsolved problem of classical physics" by Richard Feynman [60], turbulence remains as one of the most relevant problems in science [61]. Turbulence or turbulent flow is characterized by chaotic changes in pressure and flow velocity, caused by excessive kinetic energy, which overcomes the damping effect of the fluid’s viscosity [62]. Turbulence is an example of a complex system, where power laws and scale invariance of fluctuations are used to recognize the evidence of its complexity [59]. In turbulence, the dimensionless Reynolds number [63], the ratio of kinetic energy to viscous damping in a fluid flow, is used to determine its onset. Richardson’s notion of turbulence was that a turbulent flow is composed of eddies of different sizes, where the energy “cascades" from larger to smaller eddies [62]. Our present basis for the theoretic understanding of turbulence refers to the original Kolmogorov theory [62], whose validity is under evaluation within the air-sea interface [64], and assumes that turbulence is statistically self-similar at different scales [62].

Causal analyses using information theoretic metrics in turbulent flow have been discussed in a few studies, primarily to infer the emergence of coherent structures, such as eddies, generated in turbulence regime [65], [66]. Cerbus and Goldburg [67] established that the information content in turbulence using entropy density, is a decreasing function of the Reynolds number (Re) as Re increases. Using Granger causality, Tissot, Lozano-Durán, Cordier, *et al.* [68] found causal relationships between different velocity components in a wall-bounded turbulent flow using high-dimensional snapshots. These studies have provided a remarkable advance in the exploration of coherent structures in turbulence from a causal perspective using flow components. However, a multivariate causal analysis of its causal structure to determine its evolutionary dynamics has not been explored before.

In this dissertation, we provide a novel view of turbulence at the biosphere-atmosphere interface from a multivariate perspective. We implement our approach for the evolution of the causal structure using consecutive non-overlapping DAGs from observed time series, each of them constructed with 3-minutes of high frequency data (i.e., 10Hz). We use time series of high frequency observations of horizontal (WS) and vertical (U_z) wind velocity, air temperature (T), and mass density of water vapor (H_2O) and CO_2. Atmospheric pressure is also measured at the tower at the same height, however this variable is not included in the analysis given the difference in accuracy of the barometric sensor ($+/-1.725kPa$) with respect to the other high frequency sensors, which creates constant values of atmospheric pressure for periods under one-second as seen in the 10Hz data. We use the previously mentioned five time series of near-instantaneous observations to understand the behavior of the system. We particularly explore how its dynamics change under a clear sky day and under the effect of an intervention on the dynamics of the whole system, which is given by a solar eclipse, as discussed in Chapters 3 and 4.

1.2 A unique network of observational sites for the study of critical interfaces

In the Midwestern US, a massive human-induced landscape transformation from prairies into agricultural fields started around 1870, when European settlers initiated modifications to the landscape for agriculture. Tile-drain installation started around 1900, in addition to hydraulic modification of ditches to drain the low-gradient landscape [69]. Under steady intensification of industrial agriculture [70], the landscape has undergone a critical transition from a transformation-dominated system characterized by long residence times of matter and energy to a transport-dominated system that rapidly mobilizes water, carbon, sediments and nutrients to water bodies that eventually leave the ecosystem [69]. Therefore, the investigation of the role of these critical interfaces for regulating the storage, transport, and transformations of material such as water, sediment, carbon, and nutrients, are of significant importance for society. Multidisciplinary research efforts to study the Midwestern US and the role of its heterogeneous structure when responding to environmental stressors is relevant given that this region has a population of more than 61 million people, about 20% of the national total, and generates a regional gross domestic product of more than $2.6 trillion (i.e. about 19% of the national total)[71]. This dissertation contributes to this effort [26], [27] by providing unique insights at the biosphere-atmosphere interface in this intensively managed landscape.

The active layer of the Earth where soil, water, and atmosphere interact that is essential to maintain life has been recognized as being of vital importance. In the Midwestern US, this critical zone experiences strong changes in composition due to human activities [69]. Changes in land cover, water availability, and carbon cycles affect vegetation growth and how ecosystems adapt to changes in climate [71]. In this context (Chapter 2), we investigate the effect of land cover heterogeneity on the observed land-atmosphere fluxes. We particularly discuss the dynamics and origin of vertical fluxes of latent and sensible heat flux, as well as for CO_2 flux, and we discuss the behavior of the agricultural landscape as a sink/source of atmospheric CO_2 when considering photosynthetic, autotrophic and heterotrophic respiration that contribute to the variability in the vertical flux of atmospheric CO_2. We consider vegetation functions, for both maize and soybean fields as an important aspect in the study of land-atmosphere interactions due to their ability to regulate the storage and redistribution of water, carbon, and nutrients in natural and agricultural systems. However, we do not assess the final fate of the atmospheric CO_2 carried in the crops after harvest, nor the stabilization, transformations, and lateral transport of carbon in the gaseous or aqueous phase. We consider some relevant agricultural management practices such as crops rotation, and the demarcation of crop fields on the landscape that creates the "patchwork quilt" of different crops that are characteristic features of the dominant mosaic of crops in the agricultural landscape of the Midwestern US [72]. Other management practices that impact biochemical's cycles are not considered in this study, such as the application of fertilizers, the effect of tillage, or the impact of machinery use [70]. At the study site, the crops are rainfed, therefore no water-use practices are considered.

1.3 Research questions and contribution

In this dissertation, the primary goal is to understand how the information carried in the observations from an eddy covariance tower can be used to infer changes in the structure, evolution, and functionality of the system at different scales using the information that is encoded in the observations. To achieve this goal, the following research questions are addressed:

1. What is the role of the human-induced “organized land cover heterogeneity” on the fluxes observed at a 25m tall eddy covariance flux tower?

   (a) We combine 2-d flux footprint analysis and ecohydrological modeling to quantify the impact on land-atmosphere fluxes due to the heterogeneous mosaic for each crop type within the flux footprint.

   (b) We estimate the relative contribution of the heterogeneous land cover in the flux observations, which results from the fraction of maize/soybean inside the dynamically changing flux footprint and hydroclimatological drivers.

2. Are there different types of causal structures embedded in the day-time turbulence of the land-atmosphere exchange?

   (a) We move beyond the traditional spectral and/or correlative analyses when studying turbulence to explore causal dependency among the components of a multivariate system at high frequency using information flow.

   (b) We develop a novel approach to find patterns of similar behavior using DAGs, which is based on a distance-based classification to characterize the structural differences between the DAGs and a k-means clustering approach that objectively selects an unbiased number of clusters.

   (c) We explore a large range of feasible dynamics of the system by using high frequency data during a clear sky day and during a solar eclipse in 2017, which allow us to discern the robustness of the causal structures.

   (d) The evolution of the causal structure approach is used to find well-defined clusters of similar structural connectivity and the emergence of patterns of similar dynamics as the system evolves. Our results show the ability to determine patterns of causal structure inherent in turbulence.

   (e) We conjecture that variables in the turbulent system coexist and inform each other in a persistent re-organizational state around a range of dynamical behavior instead of being in a fixed configuration.

   (f) We define the “system resilience to information flow", which reflects the ability of a system to continue functioning when there has been a disruption or intervention in the typical behavior of the system, such as

under the effect of a solar eclipse, which results in a rapid adaptation of the system to changing incident solar radiation.

3. Does the evolution of the causal structure reflect different underlying dynamics? Based on the structural causal model of turbulence, are functional differences the reflection of causal structural differences, or do different causal structures give rise to the same functional behaviors?

   (a) We discern the structure-function duality in the causal interactions that shape the observed dynamics of our complex system. In other words, we determine how the causal structure of the dynamics of an ecohydrological system is related to its functionality.

   (b) Using the DAGs resulting from the clustering analysis of the evolution of the system's causal structure, we examine the differences in their information flow.

   (c) We determine that functional differences in the information flow of the system are a reflection of causal structural differences, promoted by the system itself to meet the information flow needs of the evolving system.

   (d) We define the relationship between the interactions in the causal history of information flow of each variable and the observed long-range dependency as exhibited through spectral analysis. We found that the cause for long-term dependencies in turbulent systems that do not follow the power law in the spectral analysis are caused by strong self-dependencies in the information flow coming from the distant causal history.

1.4 Organization

The Chapters of this dissertation are organized as follows:

- In Chapter 2 we explore how the influence of the heterogeneous land cover is encoded in the flux data. Flux tower measurements quantify the magnitude and temporal variability but also carry the influence of spatial heterogeneity within the flux footprint. We characterize the effects of this heterogeneity on latent heat (LE), sensible heat (H), and CO_2 fluxes (F_c) using a combined flux footprint and ecohydrological modeling approach. This helps us understand how the observed flux magnitudes and variability depend on the structure of the land cover inside the flux footprint.

- In Chapter 3 we explore the evolution of the causal structure created by the complex system of interactions in high frequency multivariate data, which contains information about the interdependency between variables such as wind, humidity, temperature, and CO_2. We move beyond the traditional spectral analyses and explore causal

dependencies among components of the system estimated using information-theoretic metrics. The resulting structural causal model is a directed acyclic graph (DAG) for time series. To estimate the evolutionary behavior of the causal structure, we develop a novel approach to find patterns of similar behavior which is based on a distance-based classification to characterize the structural differences between the DAGs and a k-means clustering approach that selects an unbiased number of clusters. This novel approach of evolution of the system structure provides a key methodological framework to find patterns in the behavior of the dynamics of a complex system and understand how the causal dependence of turbulence in a multivariate framework manifests in the high frequency data.

- In Chapter 4 we build upon of our structural causal model that arises in the form of a directed acyclic graph (DAG), to explore the relationship between the causal structure and the associated functionality of the system. The functional comparison of the sequences of DAGs shows how the information flow changes in time, possibly as a result of structural differences. Therefore, we determine if functional differences are a reflection of structural differences, or in spite of structural differences, there are no functional differences. The approach implemented here for investigating the functionality of the system through the evolution of information flow, supported by the novel approach of the evolution of the system causal structure, provides a methodological framework to understand how this complex system evolves to determine behavioral patterns of its dynamics.

- Chapter 5 summarizes the key findings and concludes the study providing avenues for future investigation.

- In the Appendix, supplementary information is provided.

Chapter 2

Inside the flux footprint: Understanding the role of organized land cover heterogeneity on land-atmospheric fluxes

2.1 Introduction

[1]

Agricultural landscapes dominate the US Midwest, influencing ecohydrological responses where the root-soil-canopy-atmosphere continuum act as an integrated system. In this region, small-grain production was replaced about a century ago by maize and soybean row crop agriculture. Today, a seasonal human-induced reorganization of vegetation to meet agricultural ecosystem services determines the spatial distribution [70], and the region experiences seasonal transitions in land cover every year. Specifically, row crop agriculture consists of seed planting in early spring, rapid growth in early summer, maturity in late summer, and harvest during autumn. During July, the US corn belt is now 40% more productive than the Amazonian rain-forest [73] as a result of steady agricultural intensification over the past two centuries. This dense vegetated land cover during the growing season contrasts drastically with an almost bare landscape of soil, roots, and litter left after harvest, typically around mid-October to November [72]. During the growing season, a patchy mosaic of different crops is the dominant feature, which partially hides other sources of heterogeneity such as soil properties and micro-topographic variability [28]. In this study, we focus on the role of the "organized land cover heterogeneity", which we define as the human-induced spatial organization of the landscape, on the fluxes exchanged between the landscape and the atmosphere when measured at a tall eddy covariance tower.

[1] This chapter is in review for Agricultural and Forest Meteorology [Hernandez Rodriguez, Goodwell, and Kumar, 2022]

We quantitatively address one of the major challenges facing the interpretation of eddy covariance measurements in heterogeneous landscapes: Besides other sources of landscape heterogeneity, what is the role of the human-induced "organized land cover heterogeneity" on the observed fluxes by a tall eddy covariance flux tower?

Eddy covariance measurements require a homogeneous flow field to provide an accurate integration of fluxes at the land-atmosphere interface [3], [6]. However, for tall towers the dynamic upwind surface area where the land-atmosphere exchange flux is generated, known as the flux footprint, generally exhibits spatial heterogeneity and fluxes from different sources mix at the observation point [14]. The use of footprint models for interpreting micrometeorological observations is a common practice, but the process of differential weighting within a temporally varying flux footprint is a "well-known but frequently overlooked feature of eddy covariance measurements" [11], [13]. Previous studies have related eddy covariance flux tower observations to individual land use, mostly using a combination of different measurement techniques at different scales. One approach relies on in situ data, from nearby towers at which flux footprints cover a specific vegetation type [22]–[24] or from flux chamber measurements [13]. However, in highly heterogeneous systems with mixed vegetation and soil wetness, it is known that there is a possibility for a serious mismatch between eddy covariance flux measurements and in-situ measurements for determining specific fluxes associated with a land cover classes. In our case, when tens of plots are located inside the several square kilometers of size dynamic flux footprint, on-site measurements might not be representative of the average behavior of each land cover type inside the tower flux footprint, which can potentially bias the conclusions of the study. Another approach relies on the use of remote sensing or aircraft data to estimate fluxes from plots on the flight transect [25]. However, the scale of heterogeneity discernible depends on the flight altitude, and if fluctuations occur during the acquisition of the data, the estimation of fluxes from different sources can be affected. Typically, the use of in-situ techniques, such as flux chambers and remote sensing including aircraft data, are limited to study cases due to their inability to provide continuous observations. To answer our research question, we require to pair the ecosystem observations from our tall flux tower, with a technique that can provide continuous data of the average behavior of each land cover type inside its flux footprint. Together, we can determine the relative contribution of each land cover type at every time step of the measurements. Therefore, instead of using scattered observations as representative of the behavior of each land cover type, we estimate the average behavior of each land cover type using an ecohydrological modeling approach. An advantage of pairing the tower observations with modeling results for each land cover type is the possibility to generalize the implementation to other land cover types in future studies. Other studies have focused on extracting the time series associated with a plot-nearby to a flux tower. In that case, the time series for the plot is obtained by extracting the fluxes observed by the tower when the plot intermittently lies within the dynamically changing flux footprint. For that purpose, TOVI software [74] can be useful, especially if the plot is located in the upwind direction from the tower. However, many times it requires additional sources of information such as nearby towers or flux chambers, to later recreate a full time series for a plot. Previous studies have used a set of towers with overlapping flux footprints or modeling results for times when the tower do not see

the area of interest [24]. Our work is distinct from these previous efforts, given that to answer our research question we combine ecohydrological modeling with a tall eddy covariance flux tower that accounts for the organized heterogeneity of the land cover and, using both, we determine the relative contribution of each land cover type to what is observed by the flux tower's long-term continuous observations. In particular, it is distinct from chu2021representativeness in that we consider a structured heterogeneity in the flux footprint whose contributions are dynamically changing at every measurement time step (15 min). chu2021representativeness consider monthly flux footprint climatologies for many Ameriflux sites, creating awareness of the spatial mismatch in model-data benchmarking. However, when aggregating over time, the flux footprint climatology blends the sources and sinks of the flux while identifying the spatial extent and temporal dynamics of the areas contributing to the observed fluxes at a tower site. To solve our research question, we take a more detailed perspective to analyze the relative contribution of each land cover type inside the dynamic flux footprint at each time step (15 min).

To measure the dynamic land-atmosphere exchange of heat, water, and carbon dioxide (CO_2) fluxes in intensively managed agricultural landscapes, such as that located in Illinois, where maize and soybean plots dominate the landscape, a 25m-tall flux tower sees hundreds of agricultural plots inside the dynamic flux footprint at each time step of the measurements [25]. The distinctive contribution of this study is to investigate how the combined action of (1) hydroclimatological drivers acting on the ecosystem, and (2) the difference in the fraction of land cover types in the flux footprint due to the organized heterogeneity of the land cover, creates the flux signal observed at the tower. Here we emphasize the role of each land cover type on the measured exchange of water, heat, and CO_2 fluxes at the land-atmosphere interface, which is a critical aspect when accounting for fluxes' sources and sinks from agricultural landscapes [75]. Given that the change in flux observed at the tower could either correlate to a change in fluxes from the crops or a shift in relative land cover contributions, we analyze the effect of the relative contribution due to maize and soybean within the footprint as a result of wind speed and direction, and atmospheric stability. Using the observations of the tall eddy covariance flux tower and other available data sources in a complementary way, such as flux footprint and ecohydrological modeling results, we are able to provide a more informed interpretation for the behavior of the observed fluxes by a tall tower with respect to the origin of the fluxes in the landscape.

In this chapter we describe the Intensively Managed Landscape Critical Zone Observatory (*IMLCZO*) study site [27], and, we present methods to account for the organized land cover heterogeneity, including considerations for the estimation of the two-dimensional flux footprint and the description of the use of the ecohydrological model to estimate the fluxes of the upwind area sources and sinks. We include the analysis of the ecosystemic flux behavior at the study site as observed by the flux tower. From the results of the flux footprint and ecohydrological modeling, we analyze the seasonal and inter-annual evolution of the flux contribution due to each crop type. Later, we connect our results to maize and soybean crop yields to investigate CO_2 flux dynamics. We end this chapter by summarizing the main findings and discuss some assumptions used in this work that could be relaxed in future studies.

2.2 Study Site

2.2.1 Description

We use hydrometeorological data and flux measurements from a 25m tall eddy covariance flux tower in the Intensively Managed Landscapes Critical Zone Observatory *(IMLCZO)*, located at 40.155N, 88.578W, Goose Creek Township, Piatt County, Illinois, US (Figure 2.1). In the Upper Sangamon River Basin both glacial and management legacies have shaped soils, topography, and native land cover, resulting in a low-relief landscape with poorly drained soils [69], [76]. Therefore, the use of tile-drains is a common practice in the crop fields for subsurface drainage [27]. The climate at the study site is humid continental (Koppen climate classification Dfa) with warm and humid summers and cold winters. Historically, maximum precipitation occurs in late spring and early summer (i.e., April to June) with an average of about 100 *mm* per month [77] and long-term observations have shown that Illinois has become wetter during the crop-growing season [77].

In this agricultural landscape, vegetation dynamics are a strong determinant of land-atmosphere fluxes and their seasonality in the landscape. These dynamics are highly influenced by crop rotation between maize and soybean fields every one or two years, which is a prevalent practice, along with different intensities of tillage [27]. The region has a return of one harvest per year. Planting occurs from early April to late May, and harvest occurs from late September to early November. Maize is typically planted before soybeans and harvested after, such that it has a longer growing season [72]. In this study, we consider an April-March window as a "crop year" (e.g. April 2016 to March 2017 is denominated in this study as "crop year 2016"). Both crops have a peak vegetation cover with a very dense leaf area index *(LAI)*, reaching 4 for maize and 6 to 7 for soybean [78]. A distinctive feature of this agricultural region is how the dense vegetation cover during the growing season contrasts drastically with the almost bare landscape left after harvest, which lasts until the following spring season when planting occurs [72] (Figure 2.2). After harvest, crop residues, i.e., mainly litter, stover and plant roots, remain on the surface and in the shallow soil layers until the following spring when planting occurs [79].

2.2.2 Instrumentation and data

Our 25m-tall eddy covariance flux tower sees the combined response of hundred of different plots every 15-min in the "patchwork quilt" landscape inside its several square kilometers size dynamic flux footprint. We use a set of detailed land cover maps [81] to characterize the annually varying spatial land cover composition. Although the underlying vegetation is non-homogeneous, the tower is situated on a terrain that is generally flat in all directions for an extended distance upwind, making the study site ideal to explore land-atmosphere flux dynamics resulting from land cover changes in a human-induced agricultural landscape. The eddy covariance tower has recorded data from April 2016

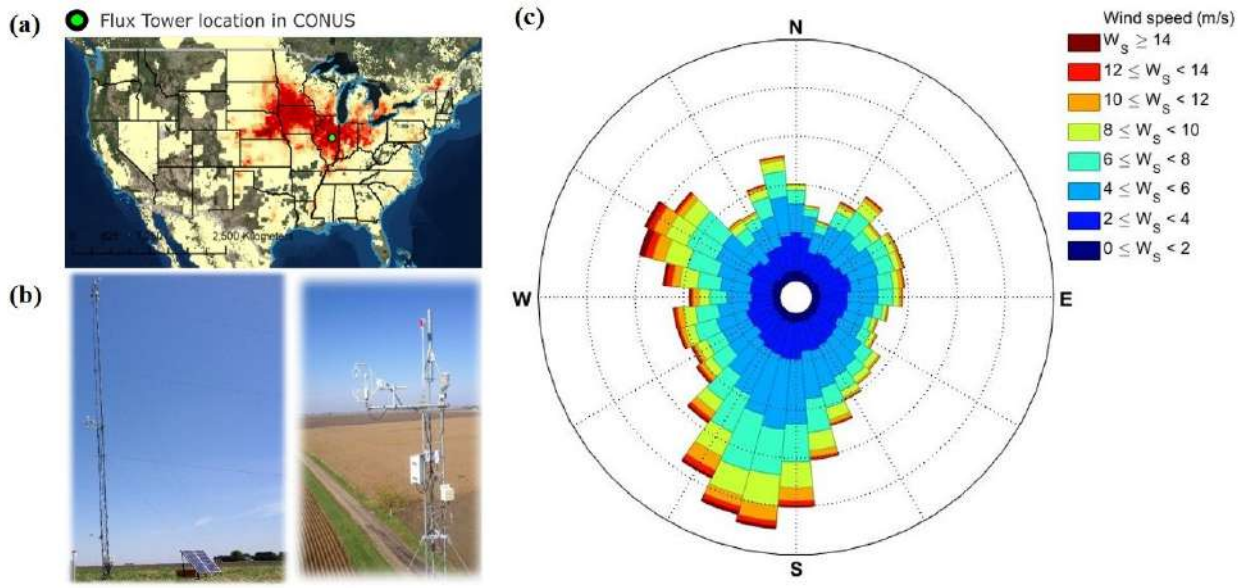

Figure 2.1: Goose Creek flux tower location and components. *(a)* Location of the tower (green dot) and the intensity of maize cultivation (the red area represents the corn harvested area fraction (high=red, low= light ivory), [80]). *(b)* Fluxes measured by the 25m tall eddy covariance tower come from the underlying heterogeneous landscape consisting of a mosaic of maize and soybean fields, from a fetch that can reach up to 10 km upwind from the tower. *(c)* The prevailing wind direction is from the southwest (4/23/2016-4/30/2019). The relative frequency with which the wind blows from a particular direction is proportional to the spoke's length, and colors indicate different wind speed categories.

to present day. The high frequency instruments that estimate fluxes from the ecosystem are deployed at 25m height (Li-7500 Infrared Gas Analyzer manufactured by LiCor Inc., and CSAT3 Sonic Anemometer manufactured by Campbell Scientific Ltd) (See appendix A.2 and Table A.2). These instruments sample at 10 Hz and are set to record 15 minute averages. They point to the southwest, the prevailing wind direction (Figure 2.1). However, constantly shifting wind directions with meteorological conditions have implications for this study (described in detail in Section 2.3).

2.3 Methods

Here we describe how we estimate the relative contribution of different land cover types to land-atmosphere fluxes measured at the flux tower. First, we use the wind data to obtain the variability of the areal coverage by using a two-dimensional flux footprint parameterization. Then we use a process-based ecohydrological model to obtain the temporally varying contribution of the flux values for different land covers. We use both the observed data and the modeled contribution of fluxes to estimate the contribution of each crop to the observed fluxes. From this, we can characterize the patterns of magnitude and variability of fluxes. Knowing that the observed fluxes at the ecosystem scale also carry the influence of the spatial heterogeneity within the flux footprint, we deconvolve the signal of the eddy covariance observation by quantifying the differential weighting of the plots based on the land cover types inside its

Figure 2.2: Panoramic view of the intensively managed agricultural study site. *(a)* During the growing season in July 2015, a row crop agriculture mosaic dominates the landscape, masking features such as micro-topographic depressions and soil variability. The vegetated land cover connects the heterogeneous ecosystem and the overlaying atmosphere during the growing season. *(b)* Right after harvest (October 2017), when only litter remains over the surface. "R" marks a common reference point between the two pictures (Photo credit: (a) Allison Goodwell, (b) Leila Hernandez Rodriguez).

dynamic flux footprint to find the relative contribution of each land cover type on the observations.

2.3.1 Estimation of two-dimensional flux footprint

Latent heat (LE), sensible heat, (H), and CO_2 fluxes (F_c) estimated by the flux tower at any given time point correspond to an uncertain origin on the landscape. This origin can be estimated as the flux footprint, which is defined as the upwind landscape area that contributes to the measured vertical flux or concentration at a specific time [6], [16], [18]. In this study, we use the two-dimensional flux footprint prediction model (*FFP*) proposed by Kljun, Calanca, Rotach, *et al.* [18], which considers the effects of surface roughness, atmospheric stability, and the crosswind spread of the footprint. For an agricultural landscape, surface roughness length changes as a function of vegetation height through the growing season. Also, atmospheric thermal stability rapidly changes with air temperature and density at a given pressure, impacting the vertical motion of air parcels. As a result, the areal contribution associated with each land cover type changes dynamically. The *FFP* model provides the width and shape of the two-dimensional flux footprint at any give time, where the source/sink area of the fluxes is located on the horizontal surface (x, y), and the tower height in the vertical direction, z (Figure 2.3). The *FFP* model assumes stationarity over the eddy-covariance integration period (here, 15-min) and horizontal homogeneity of the flow, but not of the scalar source/sink distribution.

When estimating the two-dimensional flux footprint, at each time interval the observed fluxes have their origin in a different combination of maize or soybean fields. To derive the source area up to a certain percent of flux contribution,

we define a set of five contours (r) that define the areas that contribute 20, 40, 60, 80, and 90% of the total flux estimated by the flux tower. At farther locations beyond $r_{90\%}$ that correspond to a contribution of 90%, the contributions tail off, so we limit our study to $r_{90\%}$ (we use $r_{\%}$ or r to represent percentage or equivalent fractional contribution, respectively). The associated fetch changes direction and length at every time step. In this context, the fetch is the distance from the tower with a specific fraction of the flux contribution. For example, the fetch for a 50% contribution ($r_{50\%}$) will be shorter than for a 90% contribution ($r_{90\%}$)[6].

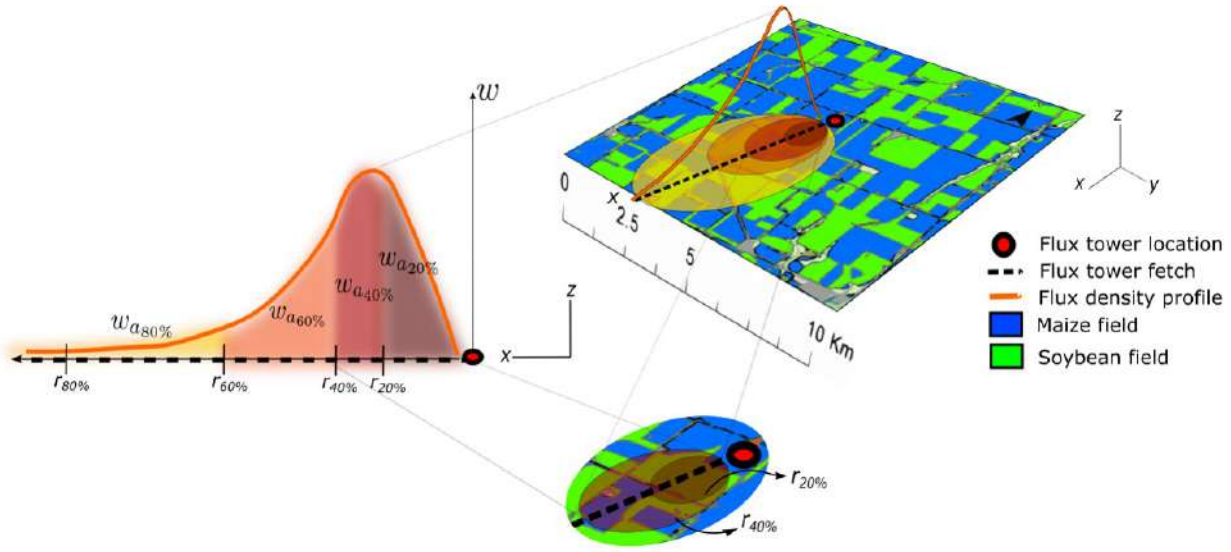

Figure 2.3: Illustration of a two-dimensional flux footprint that captures the organized heterogeneity of the maize-soybean mosaic. The land cover data from *USDA* 2016 (right) shows a mosaic of maize and soybean surrounding the flux tower. The density profile over the mosaic (orange) represents the relative contribution of the flux footprint as a function of the upwind distance, denominated as the fetch (black dashed line). The density profile shows that higher contributions come from locations close to the tower, but not immediately underneath. In the two-dimensional approach [18] the area defined by a set of contours ($r_{\%}$) of increased percentage of contribution (bottom) define the strength and location of the sources/sink areas that contribute to the flux estimated at the tower. w is the weighted flux footprint contribution of each patch of area (a) defined by a given contour. Therefore, the flux tower measurement is the combined response of the fields inside the flux footprint (left).

We used the *FFP* model as a function on a loop in our Python code to estimate flux footprints for each 15-min data point from April 2016 to April 2019. Here we describe the inputs required for the *FFP* model. The calculation of the boundary layer height, blh, is based on the bulk Richardson number, R_i, method [82] which is suitable for convective and stable boundary layer conditions and has been used in several previous studies [83], [84]. We used the blh retrieved from the fifth generation reanalysis dataset for the global climate and weather, ERA5 [85] from the European Centre for Medium-Range Weather Forecasts (ECMWF). Near-surface atmospheric turbulence is caused by thermal and mechanical effects. Thermal turbulence is produced by temperature gradients and buoyant forces, while the mechanical turbulence is generated by friction forces driven by wind shear, and therefore both control atmospheric

fluxes. To account for atmospheric stability we calculate the Obukhov length, L, [86], which is positive for stable and negative for unstable atmospheric stratification, and becomes near-infinite in the limit of neutral stratification. Finally, the last input to the *FFP* model corresponds to the standard deviation of the lateral velocity fluctuations, σ_v, which is estimated using the 15-minute root-mean-square of the cross main-wind component, v, from the high frequency data at the flux tower. The displacement height, d, is defined as the distance above the ground at which a non-vegetated surface should be placed to provide a logarithmic wind field equal to the observed one [87]. We estimated the variation of d during the growing season as a function of the average height of maize, $d = 0.67 * h_m$ [88]. The variation of canopy height and LAI has been proven to be nearly simultaneous in crops like maize [89], [90], therefore we used LAI as a proxy to define the average change of height of the crops in the region. The measurement height above the displacement height, z_m, is calculated as $z_m = z - d$, where $z = 25m$ is the tower height and d is the displacement height.

2.3.2 Heterogeneity and flux partitioning equations

Here we describe the approach to estimate the relative flux contribution due to heterogeneous land cover. Analytically, the distribution of a diffusive quantity in the lower layer of the atmospheric boundary layer is described as an integral diffusion equation. Therefore, the flux footprint relates the vertical eddy flux η from a flux tower located at the origin $(0, 0)$ and with an observation height, z_{obs}, to the spatial distribution of ground source (or sink) fluxes $\mathscr{F}(x, y)$ at the ground ($z = 0$) at a upwind distance (x) and crosswind (y) direction from the tower location [16], [17], [91]–[93]:

$$\eta(0, 0, z_{obs}) = \int_{\mathscr{R}} \mathscr{F}(x, y) \cdot \omega(x, y) dx dy \tag{2.1}$$

where $\mathscr{R}$ denotes the flux footprint and $\omega(x, y)$ is the relative contribution to the flux at any location (x, y) [18]. $\mathscr{F}(x, y)$ is the source (or sink) flux from the surface at location (x, y), with the same units as η, where $\eta = \eta(0, 0, z_{obs})$. The observed flux η is then the weighted integration of all the surface fluxes inside the contour $r_{90\%}$ of flux contribution, and has units of W/m^2 for latent heat flux, LE, and sensible heat flux, H, and $\mu mol/m^2 s$ for CO_2 flux, F_c. This approach assumes that the turbulent flow field is horizontally homogeneous [92]. Consequently, the relative contribution of each field source or sink is a function of its location within the flux footprint.

We assume that the fluxes are the same within a given patch of a crop, and there are n landscape patches that contribute to the measured flux inside the flux footprint. To account for a measure of surface heterogeneity, the source emission or sink rate for n different sources (or sinks) Equation (A.2) is expressed as:

$$\eta = \sum_{i=1}^{n} F_i \cdot w_i \tag{2.2}$$

where F_i is the ground level flux for patch i and w_i is the weighted flux footprint contribution of each patch of area a_i

located inside the contour described by any given fraction of contribution from a defined set of contours r. We can account separately for the weighted flux footprint from maize and soybean fields, represented by the subscripts m and s, inside each contour, respectively, as follows:

$$w_{m_i} = A_m \cdot r \quad \text{and} \quad w_{s_i} = A_s \cdot r \tag{2.3}$$

where the total fraction of the area covered by maize and soybean inside a contour can be denoted by A_m and A_s, respectively, as follows:

$$A_m = \frac{\sum_{\forall i \in m} a_i}{a_T} \quad , \quad A_s = \frac{\sum_{\forall i \in s} a_i}{a_T} \quad , \text{ and } \quad a_T = \sum_i a_i \tag{2.4}$$

where we can aggregate the patch areas a_i by adding the contributions of each patch type and the total area is a_T. Here we further assume that the fluxes are the same for all patches of a given crop type, such that we can aggregate those patches to compute the relative contribution of each crop/land cover type by summing the contributions as follows:

$$\phi_m = \sum_{\forall i \in m} w_i \quad \text{and} \quad \phi_s = \sum_{\forall i \in s} w_i \tag{2.5}$$

where the subscripts m and s refer to maize and soybean, respectively. We now assume that one "crop" flux value for all maize as C_m, and for all soybean as C_s. This constitutes the assumption that the vegetation type is the dominant source of flux variability, and ignores influences such as soil moisture or nutrients along the fetch. This assumption could be relaxed if detailed characteristics of each landscape patch were available. As described in the next subsection, we use a multilayer canopy model to estimate the vegetation-level fluxes (LE, H, F_c) for each crop type. Since these fluxes are modeled, here denoted as C_m and C_s for maize and soybean, respectively, their application to Equation (A.3) results in a total "modeled" flux η_{mod}:

$$\eta_{mod} = C_m \phi_m + C_s \phi_s \tag{2.6}$$

Due to these simplifying assumptions, model error, and observation errors, we expect η_{mod} to be different from the observed tower flux, η. Therefore, our aim is not to validate our modeled results with flux observations, but to merge flux tower observations, and models to improve the crop-specific estimates. Lastly, we compute the partitions of the fluxes observed at the tower for maize and soybean fields, respectively, as follows:

$$\eta_m = \eta \cdot \frac{C_m \cdot \phi_m}{\eta_{mod}} = \eta \cdot \frac{C_m \cdot \phi_m}{C_m \cdot \phi_m + C_s \cdot \phi_s} \tag{2.7}$$

$$\eta_s = \eta \cdot \frac{C_s \cdot \phi_s}{\eta_{mod}} = \eta \cdot \frac{C_s \cdot \phi_s}{C_m \cdot \phi_m + C_s \cdot \phi_s}. \tag{2.8}$$

Based on this, we combine the relative fluxes from a process-based model with a two-dimensional footprint model to determine both the fractional and actual contributions to the flux.

2.3.3 Ecohydrological modeling to simulate maize and soybean behavior

We use an annual land cover product from the United States Department of Agriculture ($USDA$) Cropland Data Layer (CDL) at 30m spatial resolution (Figure 2.3) to determine the land cover in each footprint. Some fields see maize-soybean rotation while others see maize-maize-soybean rotation. We assume that the land cover for a year does not change until the planting the next year, and therefore define a single land cover from April through March of the next year, a "crop year". We use Equation (2.5) to compute the flux footprint contribution for maize and soybean, ϕ_m and ϕ_s respectively.

We use the well-tested and validated Multi-Layer Canopy model, MLCan [78], [94], to simulate the flux response of maize and soybean, C_m and C_s respectively, under observed atmospheric drivers. MLCan uses a multilayer discretization of the canopy and root zone, including a litter layer on the soil surface, to simulate the below- and above-ground ecohydrological processes of different vegetation types. At the leaf scale, ecophysiological (photosynthesis and stomatal conductance) and physical (leaf-boundary layer conductance and energy balance) components are coupled to determine flux densities of CO_2, and latent and sensible heat, and then integrated to the canopy scale. For an extended description of the model and its validation for maize and soybean, we refer to Drewry, Kumar, Long, *et al.* [78].

We compute the fluxes associated with maize and soybean, which use C4 and C3 photosynthetic pathways, respectively. The model is driven by above-canopy measurements of air temperature $(Ta)[^\circ C]$, barometric pressure $(Pa)[kPa]$, global solar radiation (i.e. incident shortwave radiation) $(Rg)[W/m^2]$, precipitation $(PPT)[mm]$, and vapor pressure deficit $(VPD)[kPa]$ from April 2016 to March 2019. Leaf Area Index (LAI) was obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS) 8-day dataset at 500m resolution [95]. Pixels around the tower that contained a single crop were selected to estimate maize and soybean LAI. Satellite LAI measurements were complemented with LAI estimates from field measurements of Normalized Difference Vegetation Index $(NDVI)$ calculated as described by Nguy-Robertson, Gitelson, Peng, *et al.* [96]. LAI is assumed to be zero during the non-growing season, when no vegetation is on the surface and only litter from the previous season's crops remain. Maize and soybean parameters for the model are provided in the appendix A.2 and Table A.2. MLCan simulations provide LE, H, and F_c at 15 minute resolution for maize (C_m) and soybean (C_s). These are then used to compute the total modeled flux and relative contributions (Equation (2.6)).

Table 2.1: Comparison between the random and organized heterogeneous mosaic cases (Figure 2.4) towards the estimation of latent heat flux, $\eta \equiv LE$. Case A refers to the "equal contribution assumption" (Figure 2.4a), whereas Case B (Figure 2.4b) & Case C are heterogeneous situations in which soybean and maize are more dominant, respectively.

Case	η(LE)	ϕ_s	ϕ_m	LE_s	LE_m	η_{mod}	$\eta - \eta_{mod}$	diff from A
	W/m^2	*frac*	*frac*	W/m^2	W/m^2	W/m^2	W/m^2	%
A	105	0.5	0.5	120	80	100	5	0
B	105	0.4	0.6	120	80	96	9	-4
C	105	0.6	0.4	120	80	104	1	4

2.3.4 Illustration of the role of organized heterogeneity

At a single time step, we consider that the total contributions of the flux footprint from maize and soybean (ϕ_m and ϕ_s respectively) are defined by the sum of their relative contributions (Figure 2.4), as expressed in Equation (2.5). As an illustrative example, we consider three cases (Figure 2.4, Table A.1) where $\phi_s = \phi_m = 0.5$ (Case A), ϕ_s=0.4 and ϕ_m=0.6 (Case B), and ϕ_s=0.6 and ϕ_m=0.4 (Case C). Case A corresponds to equal contributions from the two crops as would be expected if the two were randomly distributed or the organized heterogeneity incidentally reflected equal contributions, like in the hypothetical case shown in Figure 2.4a. Cases B and C reflect relatively larger contributions from maize or soybean, respectively. For all three cases, assume we know that the modeled fluxes are LE_s=120 W/m^2 and LE_m=80 W/m^2, and the tower observed flux is η = 105 W/m^2. For Case A this leads to a total estimated flux n_{mod} = 100 W/m^2 (= (.5)120 + (.5)80). For Cases B and C, we would estimate n_{mod} as 96 W/m^2 and 104 W/m^2, respectively. Since for all three cases, the observed flux at the tower is the same, we can estimate the difference between η and η_{mod} (Table A.1). In this example, we see that the landscape heterogeneity of Case C leads to the closest estimate to the flux measured at the tower. We can also consider the percent difference in η_{mod} relative to the equal contributions of Case A. Here, we see that Case B leads to a -4% difference in *LE*, due to the higher contribution from maize, which has a lower *LE*. The opposite occurs for Case C (Table A.1).

This example also demonstrates that the change in the relative contribution of fluxes from two crop species can either increase or decrease the total flux observed at the tower (η) relative to the hypothetical case of a random distribution of crops. In other words, an incremental change in flux observed at a tower could either correlate to a change in flux from the crops or a shift in relative land cover contributions. This has significant implications for flux tower data interpretation in a heterogeneous landscape.

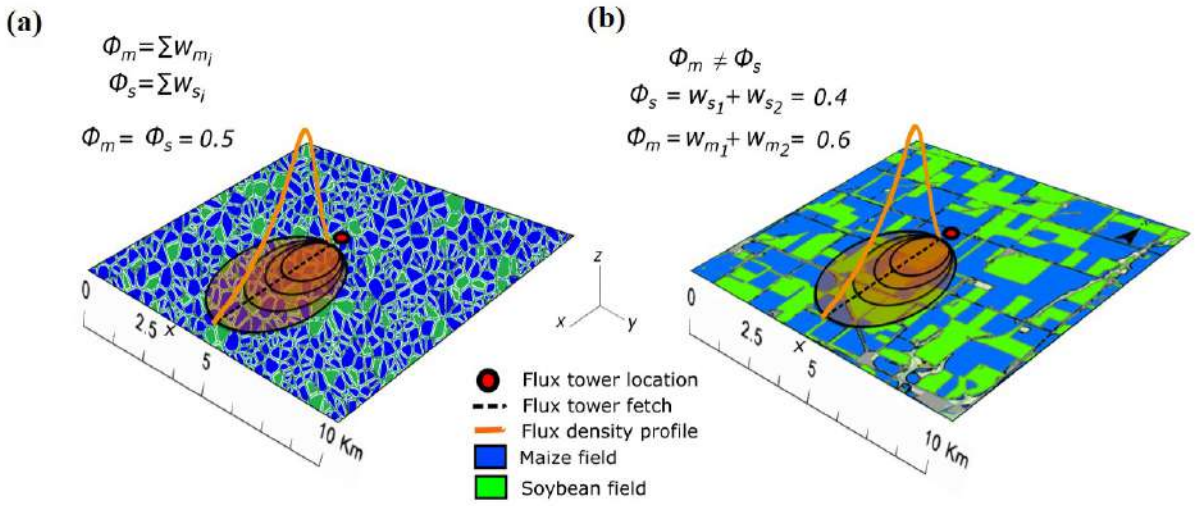

Figure 2.4: Conceptual illustration of the flux footprint responses of (a) a random land cover mosaic where maize and soybean equally contributes to the total observed flux (ϕ_m=ϕ_s=0.5), and (b) an organized heterogeneous land cover mosaic as observed at the study site.

2.4 Results and discussion

2.4.1 Ecosystem fluxes dictated by agricultural management

To understand the general behavior of the study site, we analyzed all 15 min data from 23 April 2016 to 15 April 2019 in terms of average diurnal ecosystem fluxes within the flux footprint. H and LE both peak at midday as expected (Figures 2.5a and b), and negative values of CO_2 flux, F_c, during the daytime reflect ecosystem uptake due to photosynthesis (Figure 2.5c). This absorbed CO_2 can be fixed by plants to form tissue, remain in the soil as litter and roots, or leave the system through lateral soil movement or harvested grain. Positive values of F_c during nighttime reflect a combination of autotrophic and heterotrophic respiration. We found that annually, the agricultural system acts as a sink for CO_2, with negative CO_2 annual budgets during a "crop year" (Table 2.2). Correspondingly, a lower mean annual CO_2 concentration was recorded for 2017 (Table 2.2). Our analysis also shows that crops evapotranspire on average 72% of the total annual ET during the growing season.

The difference in the net F_c (CO_2 released minus CO_2 uptaked by the ecosystem) between the three crop years (Table 2.2) is assumed to be primarily influenced by two sets of factors: (1) hydroclimatological drivers acting on the ecosystem, such as precipitation (PPT) and radiation; and (2) the difference in the fraction of maize and soybean in the flux footprint due to the organized heterogeneity of the ecosystem. Here we evaluate the former factor, while the latter, which is the main focus of this study, is discussed in the following subsections. Specifically, in Subsection 4.5, we discuss the interpretation of the behavior of the observed net F_c and its relationship with the contribution of each land cover type. For the first factor, we examined the cumulative rainfall during the crop growth period (Figure 2.5g), which

shows differences across the three years. The lowest accumulated precipitation is observed for the 2017 growing season, followed by 2018 and 2016. This reveals an opposing relationship between cumulative growing season rainfall and annual CO_2 uptake, since the 2017 growing season corresponds to the highest CO_2 uptake (Table 2.2). Initially, these results seem to contradict the notion that increased precipitation helps carbon uptake [97]. Therefore, we investigate the relationship between PPT, LE, and CO_2 flux (F_c) at the intra-annual basis (Figure 2.6c). During non-growing season months net F_c is positive, meaning more CO_2 is released rather than taken up by the system. Moreover, right after harvest, CO_2 is no longer taken up by crops and a release of CO_2 is expected from crop residues due to the organic matter decomposition of stover and roots that remains in the field [98]. On the other hand, more CO_2 is taken up during the growing season than released to the atmosphere and therefore F_c is negative. During non-growing season, LE corresponds to soil evaporation, and F_c corresponds to respiration, both of which are very low. For this system, the average monthly LE is limited to an upper threshold of about 100 W/m^2, controlled by solar radiation (Figure 2.6c). However, we observe a difference in the diurnal behavior of the inter-annual variability of the Bowen ratio (H/LE) (Figure 2.5f). For the years of analysis, the Bowen ratio is less than 1, indicating that the system is not water limited at the annual scale.

Since vegetation growth stages heavily influence land-atmosphere exchange dynamics, we explored the intra-annual evolution of the daily average behavior using 6-week windows for LE, H (Figure 2.6a, blue and orange lines, respectively) and F_c (Figure 2.6b). The intra-annual dynamics of the net CO_2 flux (Figure 2.6c) shows the windows when CO_2 uptake (negative value) is prevalent. We observe the coupled behavior of F_c and LE, for when CO_2 uptake higher LE also peaks, which means that water is evapotranspired at a higher rate. During the period of this analysis, the highest CO_2 uptake is observed from July 16 to August 31, 2017 (Figure 2.6c, window 11), the time of the year when crop growth peaks. Soil respiration might be positively related to precipitation and soil temperature [99], [100]. Therefore, a lower cumulative rainfall in 2017 may have contributed to a lower soil respiration rate in comparison with 2016 and 2018 (e.g. Figure 2.6c, shows lower CO_2 release in window 12 than in windows 4 and 20). Consequently, both a high CO_2 uptake during the growing season and low CO_2 release during the non-growing season contributed to the high net CO_2 uptake in 2017. As observed at seasonal time-scales, the non-stationary behavior of LE and CO_2 fluxes reveal a human-induced trend in which the imposed land cover controls the energy and CO_2 exchange between the land surface and the atmosphere. Consequently, the effect of vegetation on F_c is relevant, and, therefore, we quantify the relative contribution of each crop type to determine the origin of the higher CO_2 uptake observed by the tower during 2017 (See Subsection 2.4.5).

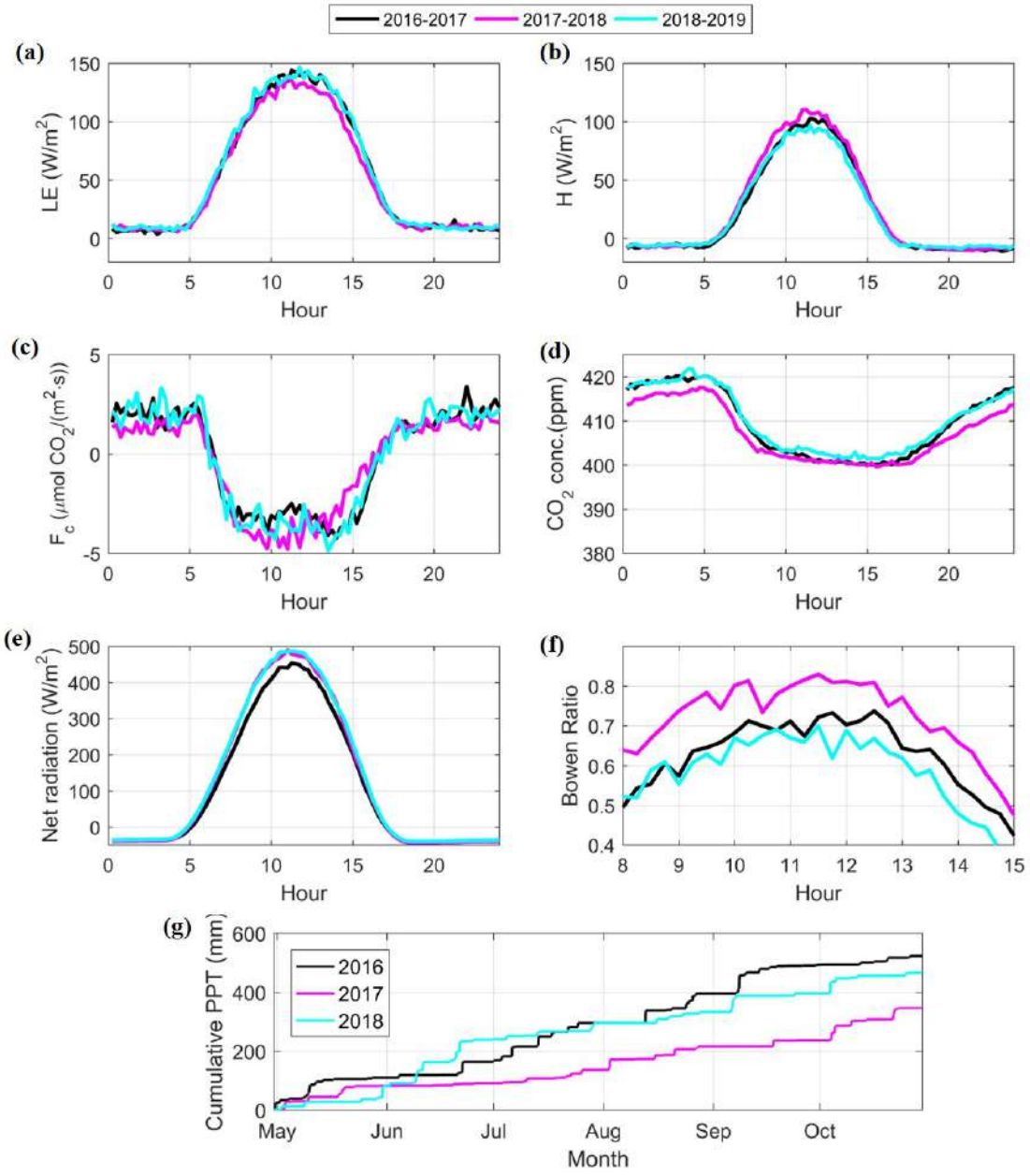

Figure 2.5: Daily averaged observations of atmospheric fluxes and radiation, and inter-annual rainfall at the study site *(a)*, for crop year 2016 (i.e. April 2016 to March 2017) (black lines), crop year 2017 (magenta lines), and crop year 2018 (cyan lines), for: (a) Latent heat flux, LE; (b) Sensible heat flux, H; (c) Net F_c (CO_2 emitted minus uptaken by the system); (d) Carbon dioxide concentration; (e) Net radiation; (f) Bowen ratio (H/LE); and (g) cumulative rainfall for the growing seasons 2016, 2017 and 2018.

Table 2.2: Annual CO_2 budget and mean CO_2 concentration inside the flux footprint

Crop year	**2016**	**2017**	**2018**
Net CO_2 budget (kg/m^2)	-0.15	-0.57	-0.23
Avg CO_2 conc (ppm)	409.16	407.02	409.84

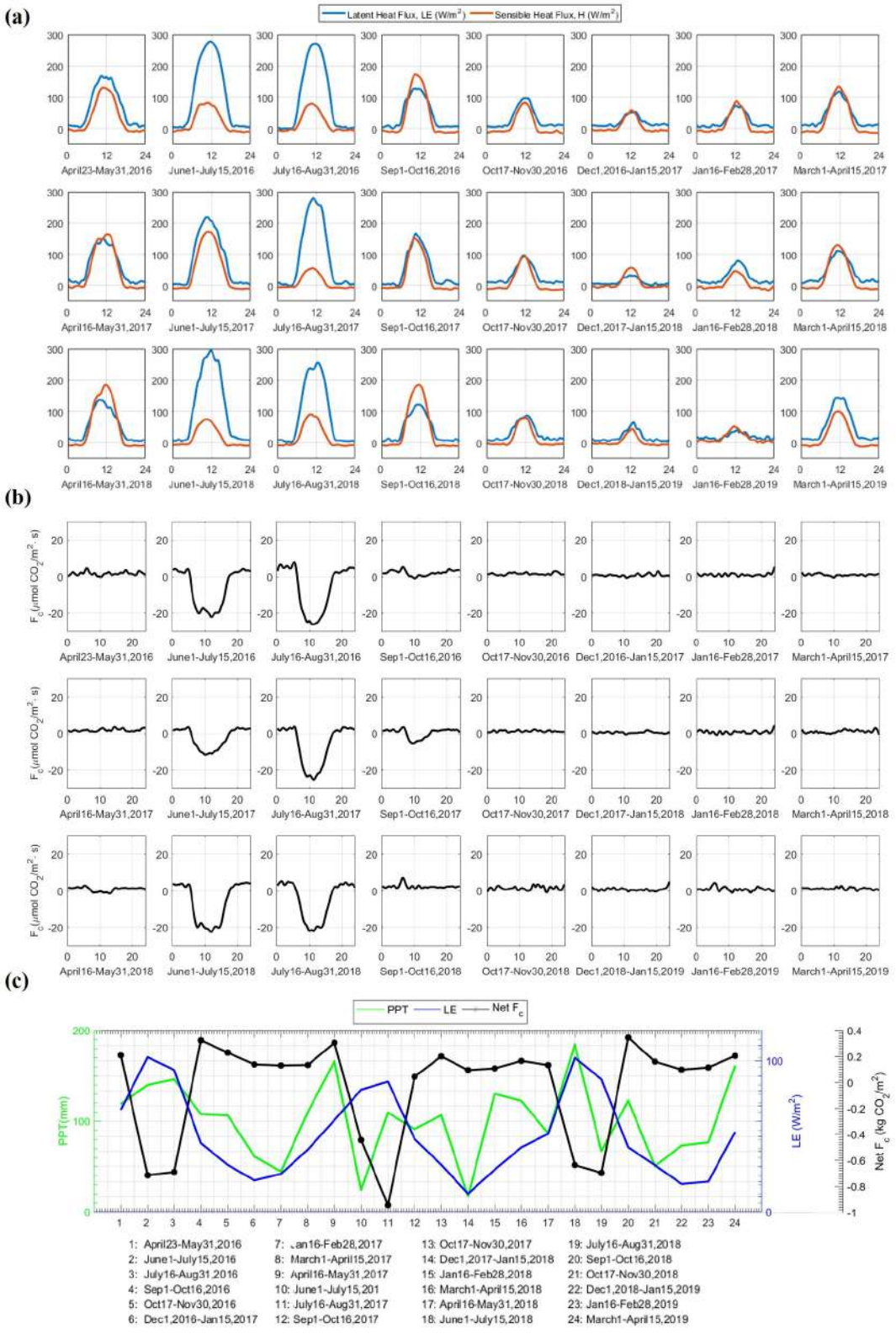

Figure 2.6: Daily average ecosystem observations at the study site in windows of 6-weeks for: (a) latent and sensible heat flux, LE and H; and (b) carbon dioxide flux, F_c. (c) Net F_c, average LE, and total precipitation PPT, for April 2016 - April 2019.

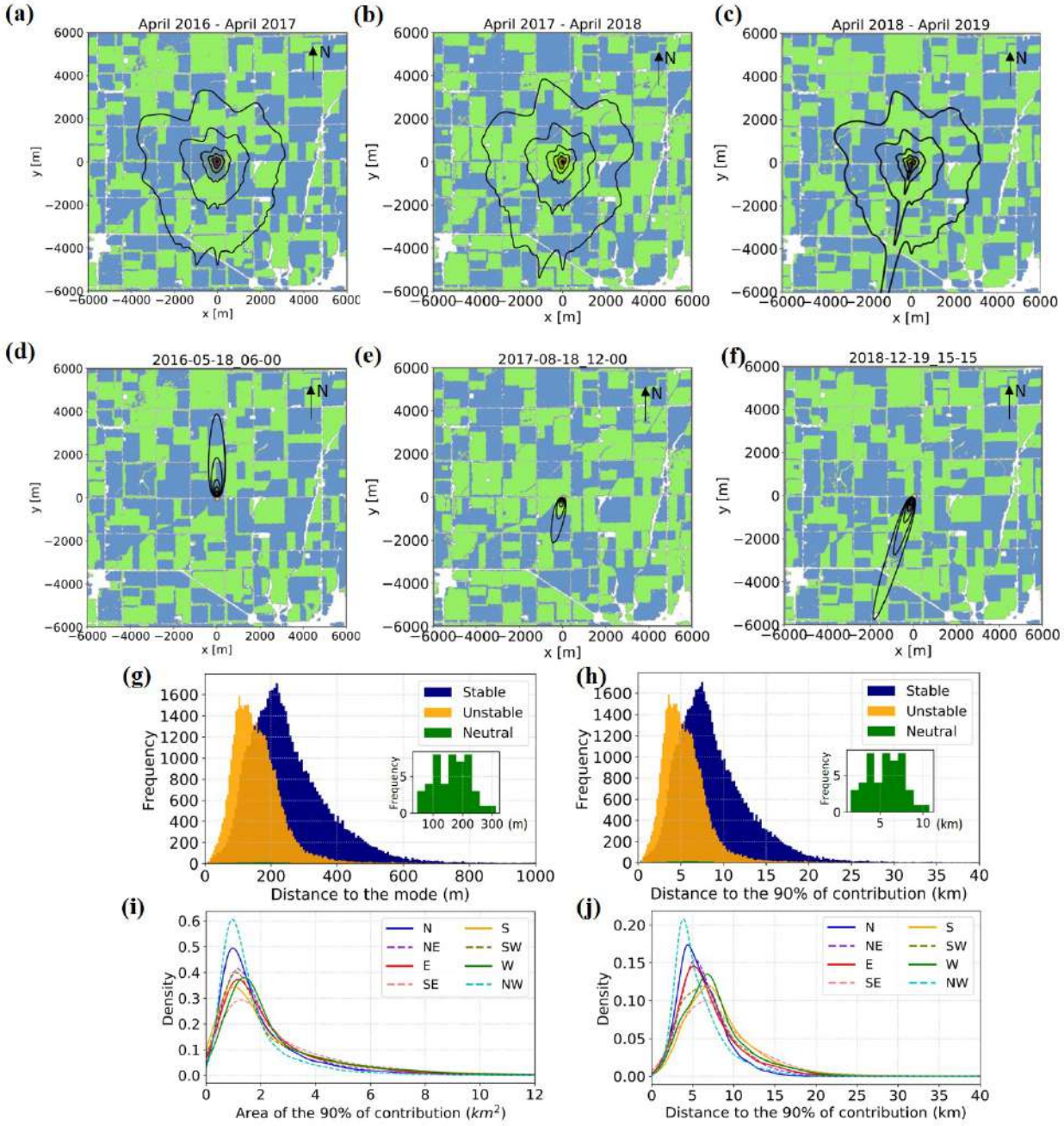

Figure 2.7: Flux footprint plots with contours over the mosaic of maize (blue) and soybean (green) crops with the center at the flux tower. Each flux footprint plot shows the source area defined by the contours of 20, 40, 60, 80, and 90% (outer contour) of the total flux contribution. The climatological (or average) flux footprint for crop years (a) 2016 (i.e. April 2016 to March 2017), (b) 2017, and (c) 2018, correspond to the aggregation of all single footprints over the year. A sample of single 15-minutes averaged flux footprints under (d) neutral (05/18/2016 06:00), (e) unstable (08/18/2017 12:00), and (f) stable (12/19/2018 15:15) atmospheric conditions show the changes in size and location of the source area of the surface flux defined by the dynamic flux footprint. Histograms of the 15-min *FFP* analysis for the upwind distance (fetch) to the (g) mode and to the (h) 90% of flux contribution, illustrate the average behavior under stable (blue), unstable (orange), and neutral (green) atmospheric conditions over the 3 years of analysis. The Probability Density Function (PDF) of the upwind (i) area and (j) distance of the 90% of flux contribution show differences based on wind direction.

2.4.2 Flux footprints cover a wide range of landscape areas

The size of the flux footprint strongly depends on the highly variable atmospheric stability at the sub-hourly time scale. At the annual scale or for observations over long periods, the effect of atmospheric stability on the footprint climatology, or average footprint, is weaker [13], [18], [101]. The changes on the climatological footprint for crop years 2016 (Figure 2.7a), 2017 (Figure 2.7c), and 2018 (Figure 2.7e) show that the mean fetch for $r_{90\%}$ is approximately 3km. A sample of flux footprints under neutral (Figure 2.7a), unstable (Figure 2.7c) and, stable (Figure 2.7e) atmospheric conditions show the different size of the flux footprint contours for each case. Turbulent mixing plays an important role in the magnitude of fields' relative contributions, as the weighting of fields farther away from the tower increases with increasing stability. In general, the flux footprint size decreases with decreasing z_m/L (Subsection 2.3.1). Therefore, under unstable conditions (Figure 2.7d) we can expect a smaller flux footprint than under stable conditions (Figure 2.7f).

We found that the average distance to the flux footprint peak is 168m for unstable conditions and 268m for stable conditions (Figure 2.7g). Therefore, the fields of peak flux contribution are located in between those distances. The footprint is wider as the standard deviation of the lateral wind fluctuations increases and, therefore, the crosswind dispersion increases [18], [101]. On average, the upwind area described by the contour of the 90% of contribution corresponds to $1.51 km^2$ for unstable conditions and $3.17 km^2$ under stable conditions. Similarly, the upwind distance to the contour described by the 90% of the flux contribution for unstable conditions is closer than for stable conditions, with 5.8km and 9.2km, respectively (Figure 2.7h). The implications of these results for the partitioning of flux contributions for each crop type due to the flux footprint are discussed in the following subsections.

2.4.3 MLCan distinguishes between maize and soybean fluxes

We used MLCan as described in Subsection 2.3.3 [94] to obtain LE, H, and F_c for maize and soybean, and to explore their intra-annual behavior (Figure 2.8). From MLCan results, we observe that LE is higher for maize than for soybean (Figure 2.8a). This occurs for all months except for June to July 2018 when LE is higher for soybean. However, at night, we observed higher condensation rates (or dew formation), in the form of negative LE, for maize relative to soybean. CO_2 uptake patterns by both crops are consistent with the observations at the flux tower (Subsection 2.4.1), with more uptake of CO_2 by the soil-crop system around noon due to photosynthesis, surpassing respiration (Figure 2.8b; negative F_c means the direction of the flux is downwards). At nighttime, positive F_c is attributed to respiration from the system. Given the dependence between LE and F_c, we explored their coupled behavior during the growing season (See appendix A.4 and Figure A.6) and we found that the four stages -planting, growing, maturing, and harvest- cluster together for maize and soybean.

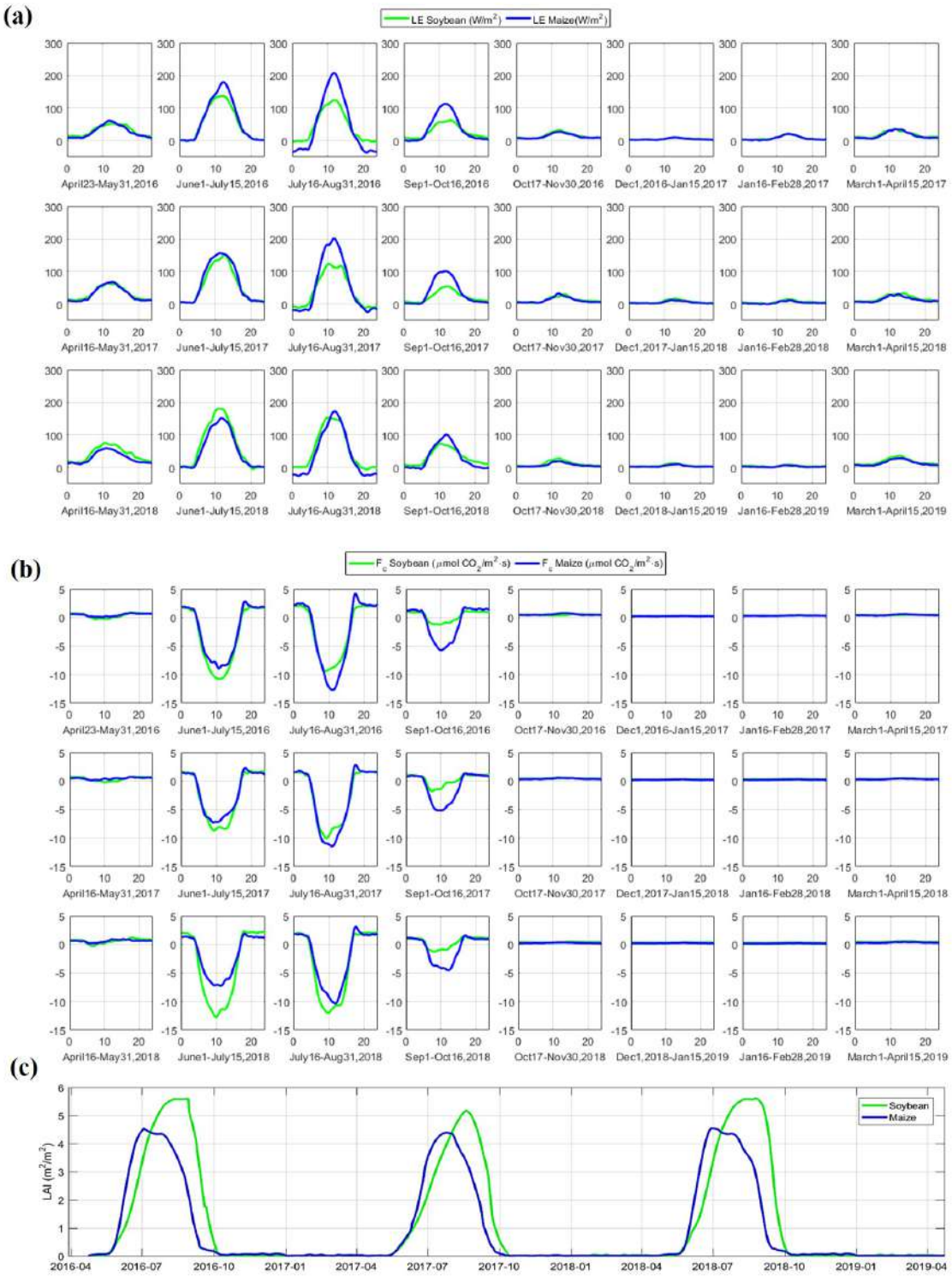

Figure 2.8: Modeled average daily fluxes of maize (blue) and soybean (green) in 6-week windows: (a) latent heat flux, LE; and (b) carbon dioxide flux, F_c, using MLCan. Model parameter values are shown in the appendix A.3 and Table A.3. (c) Canopy leaf area index (LAI) of maize and soybean inside the tower flux footprint for April 2016 - April 2019, estimated as described in Subsection 2.3.3.

2.4.4 Flux contributions evolve due to organized heterogeneity and hydroclimatological drivers

We estimated the relative contribution of average daily fluxes from maize and soybean (Figure 2.9) using the observed flux tower data. The overall trend of LE, H and F_c exhibit a strong seasonality (Figures 2.9a, b, and c, respectively). In terms of flux footprint contribution, ϕ, (Figure 2.9d), a larger contribution from maize fields were observed in 2016 and 2018, when more fields were cultivated with maize in areas between 168m and 268m upwind from the tower. In 2017, soybean was mostly planted in the same fields. However, both crop types influence the magnitude of the observed fluxes. In terms of F_c behavior across seasons, we observe that a strong release of CO_2 flux into the atmosphere occurs during planting (i.e., starting around mid April) (Figure 2.9c, numerals 1, 9 and 17) that can be explained by rising temperatures and soil moisture that support heterotrophic respiration of existing biomass on and in the soil. During the peak of the growing season, soybean fields in 2017 contributed more towards a higher CO_2 uptake (Figure 2.9c, numeral 11), while maize fields in 2016 and 2018 contributed more towards a higher LE (Figure 2.9c, numeral 2 and 18). Accordingly, this 6-week window analysis shows that maize fields inside the flux footprint contributed more towards daily LE and H than soybean fields during the 2016 and 2018 growing seasons, whereas soybean contributed more than maize in 2017. Therefore, all years have significant contributions from both crops given that the mosaic at a larger scale is different every year (Figure 2.9).

Next, we explore the seasonal dominance of a given crop type in terms of flux magnitude contribution, ϕ, rather than area contribution. We calculated the difference of the percentage of contribution for the two cases described in Subsection A.1.3. First, for a hypothetical randomly distributed mix of plants, in a manner that does not reflect any particular spatial pattern, where the relative contribution of the flux footprint for each crop type is equal, $\phi_m = \phi_s = 50\%$. Then, for the relative contribution of each crop from April 2016 to April 2019, where ϕ_m and ϕ_s are unequal and temporally variable (Figure 2.9). In other words, we want to find what crop is dominant in each window resulting from a combination of fluxes associated with each crop and the fraction they occupy in the flux footprint. To estimate the contribution percentage of each crop, we first calculated the average LE, H, and F_c for maize and soybean for the 6-week windows, using Equations (2.6)-(A.9) and the corresponding ϕ for each case. Then, we subtracted the results assuming a random (rather than organized) distribution of crops from the heterogeneous results (i.e., organized heterogeneity), to define the dominant crop for each window. The 6-week averaged analysis shows the difference between the percentage of contributions for the organized heterogeneous mosaic and for the hypothetical random contribution case (Figure 2.10). While maize fields contributed more to the observed LE and H in crop years 2016 and 2018, soybean fields contributed more in crop year 2017 (Figures 2.10a and b). The largest effect due to land cover heterogeneity is observed for CO_2 flux (F_c) from June 1 to July 15, 2017, (Figure 2.10c, bar 10), when on an average soybean fields contribution towards F_c is 24.5% larger than the random case (Figure 2.10c).

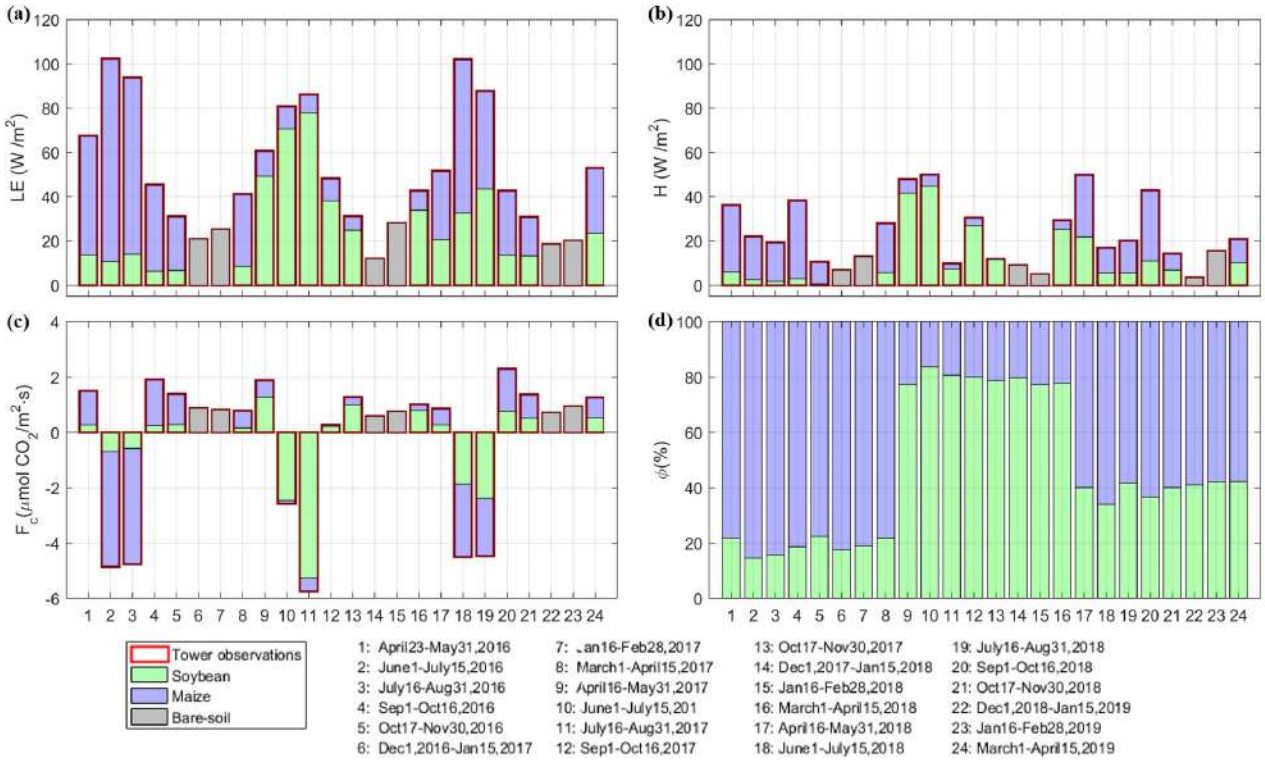

Figure 2.9: Seasonal evolution of average daily fluxes due to the relative contribution of maize (blue) and soybean (green) in 6-week windows (corresponding to the windows shown in Figure 2.6), for (a) latent heat flux, LE; (b) sensible heat flux, H; (c) carbon dioxide flux, F_c; and (d) relative flux footprint contribution, ϕ. Each stacked bar refers to 6-weeks averaged daily flux for each crop type. Relative contributions due to soybean and maize were calculated with the method described in Section 2.3 for 15-minute data from April 2016 to April 2019.

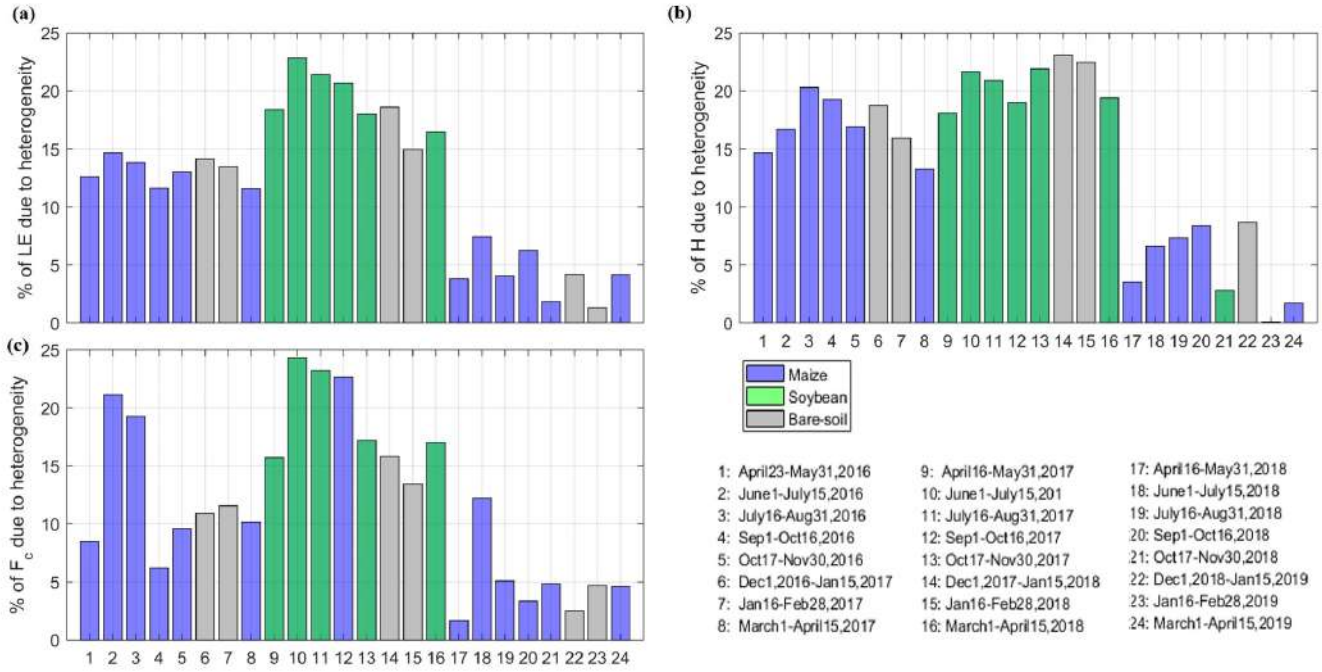

Figure 2.10: Seasonal evolution of the percent contribution due to organized heterogeneity in 6-week windows (corresponding to Figure 2.6), for (a) latent heat flux, LE; (b) sensible heat flux, H; and (c) carbon dioxide flux, F_c. These plots show the difference between the response of a heterogeneous land cover mosaic and a hypothetical random assumption where $\phi_m = \phi_s = 0.5$.

Given that the change in flux observed at the tower could either correlate to a change in flux from the crops or a shift in relative land cover contributions, we analyze the effect of the relative contribution due to maize and soybean based on atmospheric conditions (Figure 2.11). Besides the overall high contribution due to maize and soybean areas nearby the tower under unstable conditions, we observe that maize and soybean show a different cumulative distribution function based on atmospheric stability. We could have estimated the contribution of individual crops to the total flux observed at the tower at each period using a statistical approach to deconvolve the contributions from different types of vegetation. However, we would require a large amount of data to not lose resolution at times when the flux tower only observes a small area of a certain field (e.g., if the wind does not blow from a specific direction), otherwise, it is not possible to get an estimation for a given field. Particularly under unstable conditions, the flux footprint is smaller, and therefore there is a higher probability to observe a single crop (Figure 2.11b). Consequently, we combine multiple sources of information to synergistically inform the flux tower observations at the ecosystem scale and to decompose the relative contribution of each of the land cover types inside the flux footprint. This combined framework can be used to study aspects about the heterogeneity of the landscape beyond what a tower could provide.

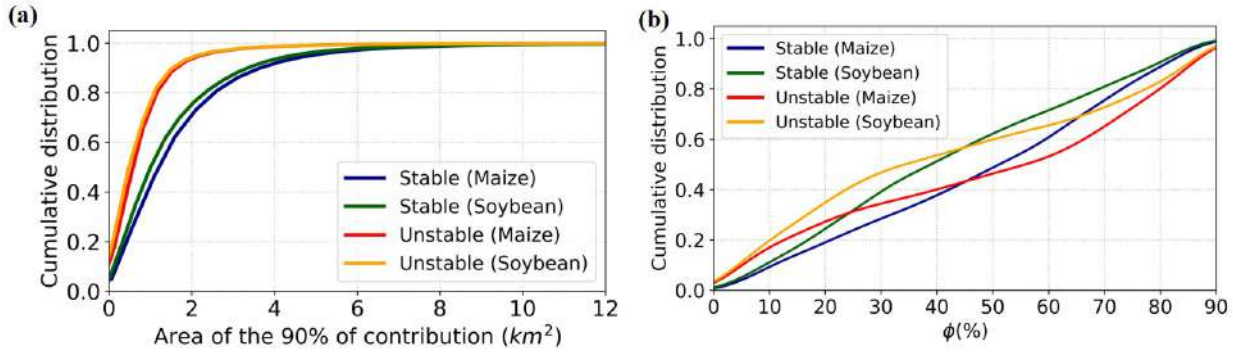

Figure 2.11: Cumulative seasonal behavior of the relative flux footprint contribution due to maize and soybean. The cumulative distribution of the (a) area of 90% of flux footprint contribution and the (b) flux footprint relative contribution, show the behavior of each crop type under stable (blue and green, for maize and soybean, respectively) and unstable (red and orange, for maize and soybean, respectively) atmospheric conditions.

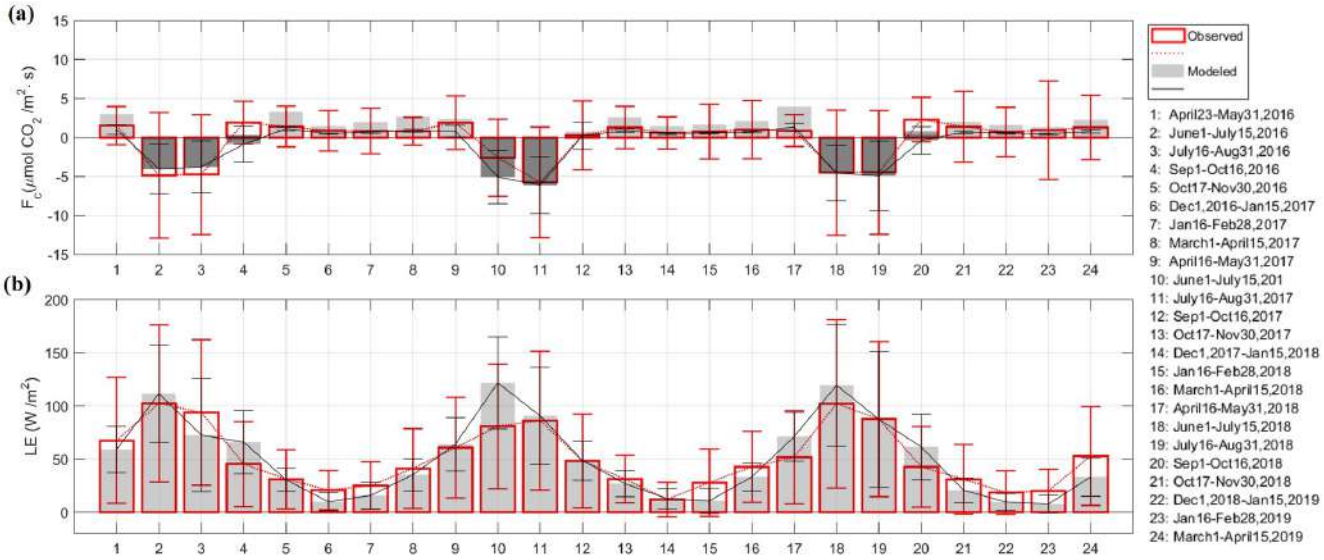

Figure 2.12: Difference between flux tower observations (red) and modeled (gray) daily averaged fluxes in 6-week windows (corresponding to Figure 2.6), for (a) carbon dioxide flux, F_c, and (b) latent heat flux, LE. The error bars show the standard deviation of the daily average over the 6-week period for the observed (red) and modeled data (black). The segmented line in red shows the behavior of the daily averaged observations, while the continuous black line similarly shows the modeled data.

2.4.5 Organized heterogeneity and hydroclimatological drivers explain the high CO_2 uptake in 2017 crop year

As mentioned in Section 2.4.1, the annual net CO_2 budget is driven by the combined action of (1) hydroclimatological drivers acting on the ecosystem and (2) the difference in the fraction of maize and soybean in the flux footprint due to the organized heterogeneity of the land cover. Here we use results from the relative contribution analysis to understand the relationship between the lowest crop yield and the largest net CO_2 budget observed in 2017, in comparison to 2016 and 2018 (Table 2.2). The $USDA$ provides crop yields at the county scale for Piatt county, Illinois [81](Figure 2.13a), which we assume is representative of yields at the study site. This low yield can be partially attributed to the high solar radiation and low rainfall during the 2017 growing season (Figure 2.5). Next, we consider how the different contributions of maize and soybean inside the flux footprint can also inform this feature. From MLCan, we obtain the net CO_2 budget for each crop type during the growing seasons (Figure 2.13b). Assuming that the CO_2 taken up is only used to produce dry matter (DM), and that the weight per bushel of DM for maize is 25.4 Kg/bushel and 27.2 Kg/bushel for soybean [102], we estimate the net CO_2 budget per unit yield for each crop type. Soybean has a larger CO_2 uptake per unit yield than maize (Figure 2.13c). From the flux footprint analysis, the relative flux footprint contribution due to soybean is larger than from maize (Figure 2.13d). Using the previous results, we estimate the net CO_2 budget inside the flux footprint for each crop type and find that the highest CO_2 uptake from soybean fields occurred in 2017 (Figure 2.13e). Given that soybean fields take up more CO_2, even in a drier, low-yield year we see more CO_2 uptake for a year in which soybean is dominant inside the footprint. Moreover, CO_2 release is muted from respiration by the drier conditions, further skewing the net flux towards high CO_2 uptake (negative Fc). Therefore, we observe that the higher annual net CO_2 budget in the crop year 2017 is not only the effect of hydroclimatological conditions but to the particular contribution of soybean fields, which play a significant role in the higher uptake of CO_2 observed that year. The combination of CO_2 taken up by soybean fields and the lower respiration rate due to high temperatures and lower rainfall, influence the overall higher uptake of CO_2 observed in 2017. These results show how the knowledge provided by multiple sources besides the observed F_c (i.e., MLCan and flux footprint modeling, as well as crop yield data), provides elements for an informed interpretation of the influence of the organized land cover heterogeneity on the behavior of observed land-atmosphere fluxes.

2.5 Summary and conclusions

This study illustrates the important role of the organized land cover heterogeneity on land-atmosphere fluxes of heat, water, and CO_2, that are measured at a tall tower. When the land cover is heterogeneous, inconsistencies in data interpretation can arise when only accounting for the vegetation type in the nearest field or assuming that multiple crop

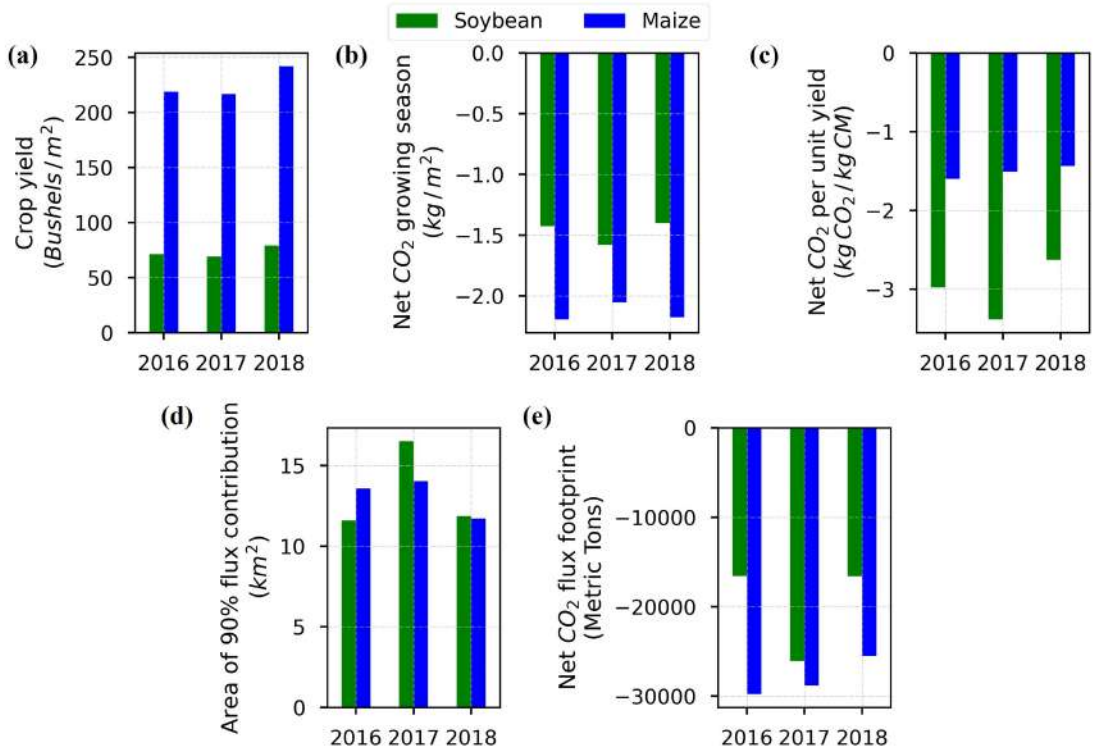

Figure 2.13: Relationship between crop yield and net CO_2 during the growing season. (a) Crop yields at Piatt county are combined with (b) the net CO_2 budget for each crop type during the growing seasons 2016-2018 estimated with MLCan, to calculate (c) the net CO_2 uptake per unit yield of dry matter (DM) for soybean and maize. (d) The climatological flux footprint area for the 90% contribution is used to calculate (e) the net CO_2 uptaked by maize and soybean fields at the study site during the growing season.

types contribute equally to the observed flux. Our framework combines flux footprint and ecohydrological modeling together with flux tower data to improve upon the understanding that could be obtained given any single information source.

Our approach of analyzing the flux contributions associated with the footprint of a tall tower footprint has some limitations, many of which could be improved upon in future studies. First, we consider that crop and soil components are the primary sources of CO_2, and that extremely low traffic in the nearby farm roads and other local sources are negligible contributors (the flux tower is strategically located away from major highways). Second, the estimation of the flux footprint is not the only source of error, but tower observations and the ecohydrological model also have errors. Third, we do not consider the variability in flux response across multiple fields cultivated with the same crop, but we assume a representative modeled flux. Specifically, we distinguish between "maize" or "soybean" patches and ignore other differentiating factors. In reality, all maize or soybean fields may not behave the same due to differences in cultivars, soil type, microtopographic variability, time of planting, etc. This assumption can be overcome through a distributed modeling approach if detailed data to support such modeling is available. For example, given a detailed map of soil texture or topography, the landscape could be divided into more than two components which could be modeled and attributed at higher resolution, e.g., "well-drained maize" or "drier soybean". This would enable testing of the assumption that vegetation type is the dominant differentiating factor between fluxes at different landscape patches. We anticipate that this study could be extended to study other natural and human-induced interventions on heterogeneous agricultural landscapes, such as varying wetness conditions, LAI, or planting dates, or for comparison with a remote sensing product. This approach could also be applied to model evaluation, in that the representation of landscape heterogeneity should lead to an improved agreement between model results and observations, relative to the assumption that the tower measurements represent a single crop type or a homogeneous contribution from multiple crop types. In general, our method is relevant for the understanding of land-atmosphere fluxes in heterogeneous landscapes and can be extended towards the use of flux tower data as validation for models of these fluxes.

Our analyzes show that the fluxes observed at the tall flux tower are the result of the combined action of (1) hydroclimatological drivers acting on the ecosystem, and (2) the difference in the fraction of maize and soybean in the flux footprint due to the organized heterogeneity of the land cover. In other words, the change in flux observed at every time of analysis at the tower (15 min) could either correlate with a change in flux from the crops or a shift in relative land cover contribution within the footprint, or both. For instance, we qualitatively demonstrated that the change in the relative contribution of fluxes from two different land cover types can either increase or decrease the total flux observed at the tower. Therefore, we quantitatively showed that the spatial structure of the land cover, described here as "organized land cover heterogeneity" and characterized by the mosaic of crop fields, impacts the observed fluxes. Our focus on the relative contribution of maize and soybean fields inside the flux footprint shows the importance of an accurate description of the land cover and the use of an accurate flux footprint method. We recognize that it is equally

important to accurately simulate the flux response of each vegetation species under the observed atmospheric drivers. All these are used to obtain the variability of the areal coverage and the temporal variability of the flux.

In an intensively managed agricultural landscape, where each land cover patch is easily identifiable by crop type (i.e. maize or soybean), we quantified the relative flux contribution of LE, H, and CO_2. At the study site, the cultivated fields where the flux contribution peaks are mainly due to one crop type, which explains the dominance in flux contribution given by maize-soybean-maize for 2016-2017-2018 crop years, respectively. This combined analysis makes it feasible to investigate questions regarding real and hypothetical land cover changes at an ecosystem scale and quantify the effects of different vegetation types on ecosystem fluxes.

Chapter 3

Causal interaction in high frequency turbulence at the biosphere-atmosphere interface: Structural behavior

3.1 Introduction

[1]

At the biosphere-atmosphere interface, fast-time-scale turbulent processes are characterized by rapid interactions between a number of variables, such as air temperature, wind speed, water vapor, and CO_2. Towers with eddy covariance instruments that are equipped with fast-response analyzers, usually at 10Hz or 20Hz, measuring turbulent fluctuations of CO_2 and H_2O molar concentrations [3], along with temperature and wind velocity components provide a tremendous opportunity to explore such interactions. These observations are typically used to estimate the vertical turbulent exchange of momentum, energy, and gases between the biosphere and the atmosphere [3], [6]. This high frequency data has been typically used for quality control, using spectral analyses to determine underlying periodicities in the behavior of the observed time-series [3], [87], [103]. While spectral approaches have provided tremendous insights into variations in single variables or co-variations among variables [104]–[107], they do not reveal the structure of causal dependencies among them. Therefore, here we propose an approach for analyzing high frequency eddy covariance measurements where causal dependencies arise as fluctuations in one variable at any time, instigating changes or variations in other linked variables at a future time [49]. The source variable affects the response variable creating a causal structure of interdependence often represented as directed acyclic graphs (DAGs) [10]. We use DAGs to identify patterns of the

[1]This chapter is in review for Journal of Geophysical Research [Hernandez Rodriguez and Kumar, 2022]

causal structure of interaction in turbulence using high frequency data (10Hz).

Previous efforts to explore complexity in the land-atmosphere interface include the use of time-series measurements of horizontal ocean-surface-wind and vertical profiles of time-height winds to investigate emergent features, such as the one detected when scanning space-time images of turbulence [108]. Using complexity measures based on information theory [109], such as mutual information and statistical complexity, evidence of regions of higher complexity have been identified where statistical models of turbulence perform poorly such as at the top of the boundary layer where transitions of stability regime occur [108]. While the location of areas of strong and weak turbulence can be inferred, how the dynamic system composed by multiple variables self-organizes to produce such gradients, and consequently, the cause of emergent patterns in turbulence remains unknown. While a decay in mutual information in ocean-surface-wind data has been observed in periods of 10Hz that can be comparable to the occurrence of gravity waves [108], drawing causal inferences requires the exploration of the dynamics of the multivariate system at turbulence scales. Recently, Krich, Runge, Miralles, *et al.* [110] estimated process graphs using the PCMCI algorithm [111] to obtain causal networks and cross-correlation-based approaches for correlation networks, to test if causal networks are considerably more informative than correlation networks. They confirmed the capacity of causal networks to extract time-lagged dependencies under realistic settings, while cautioning against the interpretations depending on the method's assumptions and datasets used to infer causality. In this study, we move one step forward in the study of turbulence to explore the evolution of the dynamic multivariate interdependencies through the causal structure of interaction.

The study of the evolution of the causal structure allows us to explore how the system self-organizes in different states of its dynamics over time. In the absence of a causal dynamic law to determine the trajectory of the dynamical states of the system, we use its causal structure to identify patterns of similar behavior and find the dynamical ranges in which the system exists. One way to ensure that we understand what are the dynamical ranges of the causal structure configurations is to perturb the system rapidly and see whether it recovers to the original state or not. In general, we assume that the biosphere-atmosphere system is an open but non-manipulable complex system. It is open given that it exchanges matter, energy, and information with its environment. Moreover, it is generally non-manipulable at the system scale, meaning that a single intervention, natural or human-induced, is generally not able to perturb the whole system dynamics. However, a solar eclipse is a natural intervention of an exogenous and relevant control on the dynamics of the whole system that allows us to explore its dynamics under the effect of the rapid decrease and recovery of the incoming solar radiation. Therefore, we explore the evolution of the causal structure of the system to determine the range of dynamics of the system for a clear sky day and under the effect of a solar eclipse, to explore the development of potentially unique causal structures of the connectivity of turbulence.

Nevertheless, defining the evolution of the causal structure of such a complex system is not trivial. Here we use information flow to characterize temporally asymmetric interdependencies among variables, as carried by the propagation of fluctuations, as the primary indicator of causal dependency [54], [55]. The causal structure then

represents the resulting network of interdependencies in a multivariate complex system characterized as DAGs [54]. Recent studies have shown that the dynamics of a complex system are shaped by multivariate interactions at different time lags [50], [51], [112], such as for stream chemistry dynamics [9] or in weather and climate [47], [55], [110], [113], [114]. However, such complex multivariate interactions in turbulence at the biosphere-atmosphere interface have not been explored before.

At high frequency, causal graphs in the form of DAGs have the potential to reveal the complexity of the dynamics, which can lead to a novel understanding of turbulence. In this study, we divide the high frequency data in a sequence of minimal size windows to contain sufficient length of data to allow for statistically robust inference of DAGs. Changes in the structure of the DAGs across these contiguous windows are then used to characterize the evolution of the causal structure as the system evolves. We use the approach developed by Runge [51] to develop the DAGs, where causal associations between subprocesses can be visualized as links in a complex interaction network. Using a distance-based classification between DAGs based on Malmi, Tatti, and Gionis [115], we characterize structural differences between pairs of DAGs to identify variations in the structures associated with the turbulent exchange. In conjunction with a k-means clustering approach, we find patterns of similar behavior in the evolution of the causal structure. In this endeavor, information theory metrics [109] play a significant role, offering a compelling approach to investigate causal dependencies between the interacting variables.

In this study, we ask: Are there different types of causal structures embedded in the day-time turbulence of the land-atmosphere exchange? To answer this question, we use the high frequency data for mid-afternoon period for two days. The first day, August 18, 2017, is a clear sky day with land-surface cover comprising of a mosaic of corn and soybean crops at their peak productivity. The second day, August 21, 2017, corresponds to a solar eclipse with maximum obscurity of 94% at the tower location. The same time windows are chosen for both days and they cover the period from onset to clearance of the eclipse. We create DAGs for all 3-minute non-overlapping but contiguous time periods for both days and k-means clustering to find common patterns in the structure. This gives us the opportunity to capture and understand the different causal structures that manifest during this period, to see if a rapid decline in solar radiation results in remarkably different causal structure. This allow us to assess if the causal structure we infer are representative of broader dynamical characteristics.

This chapter is organized as follows: In Section 3.2, we briefly introduce the study site and recall the basic concepts of information theory and DAGs. Also, we present our novel approach to estimate the evolution of the causal structure using DAGs. In Section 3.3, we present the results of the evolution of the causal structure. Section 3.4 discusses the findings on the causal structure of the dynamics of the system at high frequency. Finally, in Section 3.5 we provide conclusions for the chapter. Appendix A.2 contains a description of the instrumentation used in this study. Appendix A.5 shows an example of how to compare a pair of DAGs using a distance metric. In appendix A.6, we test the performance of the distance metric used for clustering. In appendix A.7, we determine the number of edges in a fully connected DAG

without contemporaneous edges, G_{full}. Finally, in appendix A.8, we determine the cross-covariance of the time series used in the analyzes to determine a wide range of feasible dependencies, both causal and correlative.

3.2 Evolution of causal structure

3.2.1 High frequency data from a flux tower in the Midwestern US

A 25m tall eddy covariance flux tower is located in the intensively managed agricultural landscape of the Midwestern US, at 40.155N, 88.578W, Goose Creek Township, Piatt County, Illinois, US, [26], [27] (Figure 3.1a). Instruments located at 25m height of the tower measure fluctuations in turbulent motion at 10 Hz, resulting in a set of multivariate high frequency time series. To measure the fluctuations in turbulent motion at the tower site, we used a Li-7500 Infrared Gas Analyzer manufactured by LiCor Inc. and a CSAT3 Sonic Anemometer by Campbell Scientific Ltd at 25m height [26], [27]. The incoming shortwave radiation $[W/m^2]$ is measured at 15-min (Figures 3.1d and 3.1e) using a CMR4 Kipp and Zonen Pyranometer located at 10m height on the same tower [26], [27]. In appendix A.2, Table A.2, we include a description of the instrumentation used in this study. We also include the repository where the data used in this study can be accessed.

To determine the causal structure of interactions, we use the high frequency time series of the horizontal wind speed, $WS[m/s]$, vertical wind speed, $U_z[m/s]$, air temperature, $T[^\circ C]$, water vapor, $H_2O[g/m^3]$, and, carbon dioxide mass density, $CO_2[kg/m^3]$. In this study, we use the data from a clear sky day, August 18, 2017 (**a**), and from an eclipse day, August 21, 2017, when the tower was on the path of 94% of totality (Figure 3.1b) and the maximum obscurity occurred at 12:19:53 (HH:MM:SS, Central Standard Time). To explore the evolution of the causal structure of the dynamics of the system, we subdivide the data in 3min non-overlapping contiguous windows (i.e., 1800 data points each) to ensure statistically reliable estimates of information-theoretic metrics used to estimate the multivariate interdependencies (Subsection 3.2.3). The high frequency data is standardized by subtracting the mean and dividing it by the standard deviation for each of the 3-min windows.

3.2.2 Estimating causal graphs from high-dimensional time series datasets

Quantifying causality was first envisioned by Wiener [116] and later formulated by Granger, anchored on observational data and statistical dependency. For two variables Z and Y, Granger [117] characterized causality on the reduction in the variance of a target variable Z_t due to the knowledge of a lagged source variable $Y_{t-\tau}$, where τ is a positive time lag. This approach accounts for the statistical dependency of one (or multiple) lagged source(s) on a target variable. It can be extended to consider the entire probability distribution using information-theoretic measures such as Transfer Entropy

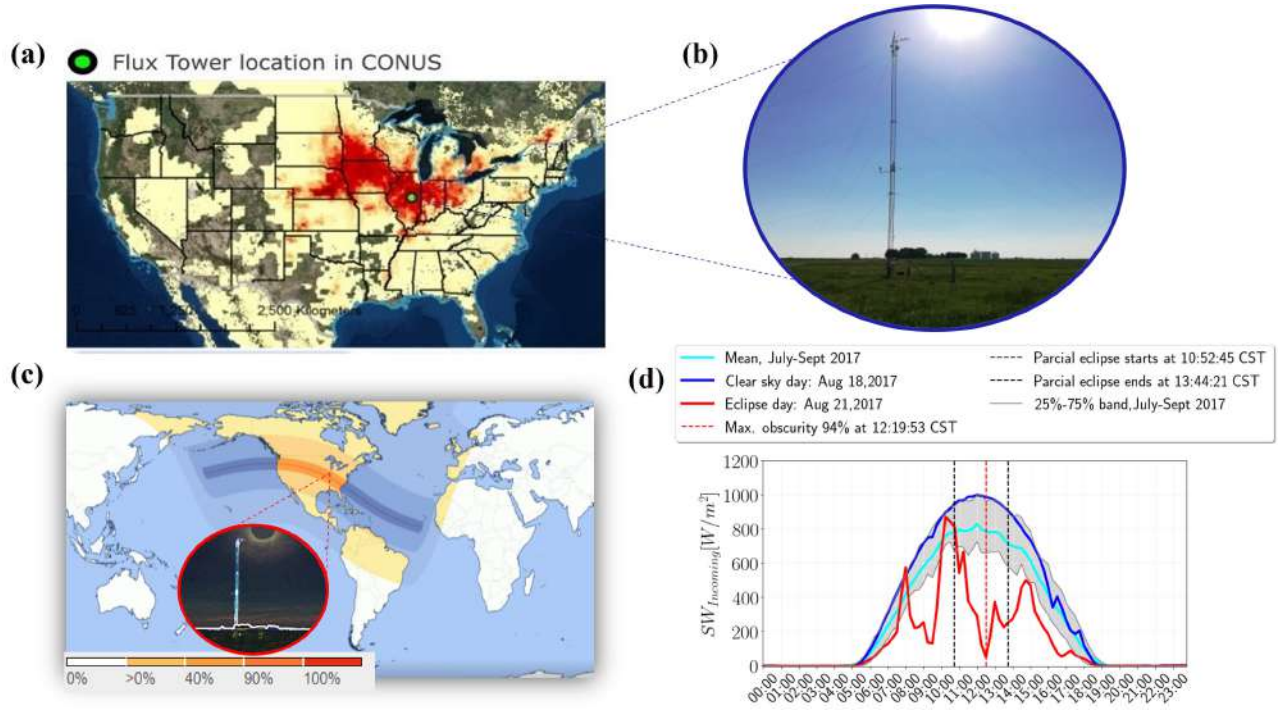

Figure 3.1: Time series of high frequency data from a 25m tall eddy covariance flux tower offers a unique opportunity to investigate the interdependencies among components of an ecohydrological multivariate system at the biosphere-atmosphere interface. (**a**) The tower is located in the agricultural Midwestern US (green dot), a region intensively cultivated with corn (high=red, low= light ivory [80]) and other annual crops. (**b**) On August 18, 2017, the tower observed a mosaic of corn and soybean crops during a clear sky day. (**c**) During August 21, 2017, the tower was under the 94% band of totality of a solar eclipse. (**d**) The attenuation of the solar radiation (red line) due to the eclipse and its contrast, with the clear sky day (blue line).

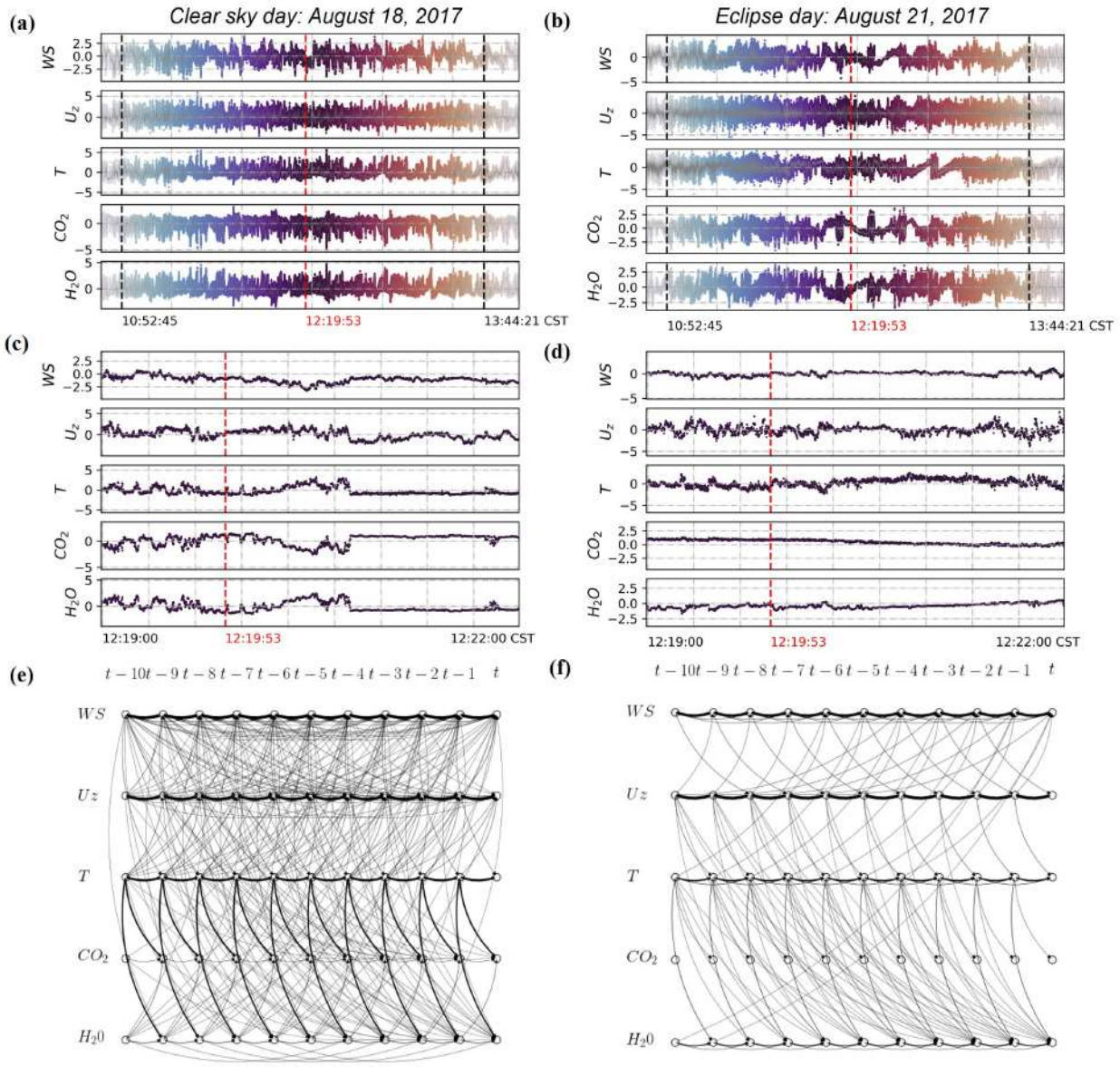

Figure 3.2: Time series of standardized 10 Hz data and the estimated directed acyclic graph representation, for five variables: Horizontal wind speed, $WS(m/s)$, vertical wind speed, $U_z(m/s)$, air temperature, $T(degC)$, carbon dioxide mass density, $CO_2(kg/m^3)$, and water vapor mass density, $H_2O(g/m^3)$. We use data from 10:46:53 to 14:01:53 CST (US Central Standard Time) for August 18, 2017 (**a**), and for an eclipse day, August 21, 2017 (**b**). The maximum obscurity (red dashed vertical line in (**b**)) occurred at 12:19:53 CST (Darker colors of the time series show proximity to the time of maximum obscurity). The partial eclipse began at 10:52:45 CST and ended at 13:44:21 CST (black dashed vertical lines). For the times series of 3min window from 12:19:00 to 12:22:00 CST for the clear sky day (**c**) and the eclipse day (**d**), we estimated the respective directed acyclic graphs ((**e**) and (**f**)) for time series using the Tigramite algorithm ([43], [51]–[53]).

[43], [118]. Regardless of whether the system can be intervened or not, information-theoretic metrics provide avenues to quantify information flow among interacting variables [56]. In this framework, the estimation of causal interactions is based on the PC algorithm [10]. Here, to estimate causal interactions we use the conditional mutual information (CMI) test based on k-nearest-neighbor estimator (CMIknn) dependency measure in Tigramite algorithm [43], [52], [119], a non-parametric conditional independence test based on a nearest neighbor estimator of conditional mutual information [119]. CMI is estimated by combining the Kozachenko-Leonenko k-nearest-neighbor entropy estimator [120] generalized to CMI [121], with a nearest-neighbor local permutation scheme. CMIknn starts without knowledge about the links and learns the set of parents, or source nodes, influencing each target variable. The links of interactions are found by testing for the conditional independence between all possible pairs of conditioned nodes and testing all combinations among them [43]. The union of edges from the parents and links without including any contemporaneous links, comprises the DAG [43]. Therefore, to estimate the joint density we make no assumptions about the parametric form of the dependencies, and the CMI, $I(X;Y \mid Z)$, between X and Y given Z, is given by [119]:

$$I(X;Y \mid Z) = \int p(z) \iint p(x,y \mid z) \log \frac{p(x,y \mid z)}{p(x \mid z) \cdot p(y \mid z)} dxdydz \tag{3.1}$$

where p represents the appropriate univariate or joint PDF. This is combined with a shuffle test that is only suitable for continuous time series and is used to generate the distribution under a null hypothesis of independence [119]. The knn estimator for the CMI is given by [119]:

$$\widehat{I}(X;Y \mid Z) = \psi(k) + \frac{1}{T} \sum_{t=1}^{T} \left[\psi\left(k_{Z,t}\right) - \psi\left(k_{XZ,t}\right) - \psi\left(k_{YZ,t}\right) \right] \tag{3.2}$$

where ψ is the Digamma function [119] and k is the number of nearest-neighbors, T is the sample length, and $k_{Z,}, k_{XZ}, k_{YZ}$ are the numbers of neighbors in the respective subspaces [119]. Large k provides a low variance of the estimated CMI, which is important for a significance test [121].

3.2.3 Directed Acyclic Graphs (DAGs) representation

A DAG for time series is a network structure denoted here as $G = (V, E)$, that consists of a set of vertices V and edges E, where an edge is directed from one vertex to another in a set of ordered pairs $(i, j) \in V \times V$. Alternatively, vertices and edges are called nodes and links, respectively. A DAG is "acyclic" in that the links never form a closed-loop [10]. In a DAG, two vertices u and v are called "connected" if G contains a path from u to v. Otherwise, the vertices are "disconnected". If the vertices are additionally connected by a path of length 1, i.e., by a single edge, the vertices are called adjacent. In this study, we use DAGs for time series showing the observed variables as nodes, such as temperature, wind speed, and water vapor, while the links denote interactions among them at lagged times (Figures 3.2e and 3.2f).

The DAG for our time series is a structure causal model [43] where it reflects a near instantaneous state of the dynamics of the system [51], whose emergence is the result of nonlinear interactions among the variables [56], [122].

The interactions among variables are seen as a result of information flow arising when fluctuations in one variable instigate fluctuations in a linked variable [43]. Directed edges between variables in a DAG are interpreted as causal links where the statistical methods used for their estimation measure the magnitude of information flow and associated time delays, while at the same time excluding the influence of other indirect influences or common drivers [43].

3.2.4 Information flow and dimensionality reduction

The estimation of information flow using a DAG for time series can be achieved through a direct edge [43], using one [52] or two causal paths [54], or through the whole causal history [9], [50]. Causal history refers to the entire set of nodes of the past in a DAG that can potentially influence the present value of any node. When accounting for information flow, the computation of CMI is infeasible due to the potential infinite number of nodes arising from the causal history (i.e., curse of dimensionality). To make the computation possible, the *momentary information transfer* (MIT) was developed by Runge, Heitzig, Petoukhov, *et al.* [43] using the Markov property for DAG [123], which states that a node conditioned on its parents, is statistically independent of the rest of the history. However, the dimensionality can still be high after using this Markov property [43], given that the complexity grows as the number of variables increases and/or as the number of lags that influence a target increases [9]. The *momentary information weighted transitive reduction* (MIWTR) technique proposed by Jiang and Kumar [9], further reduces the dimensionality of a DAG when computing the information flow using a causal history approach [9], [50], [112]. The cardinality reduction based on MIWTR does not affect the estimation of information-theoretic measures significantly when the length of data in the time series is sufficient [9]. A data length greater than 1000 data points per variable for the estimation of a DAG is considered sufficient to achieve reliable estimations of information flow [9]. We refer the reader to Jiang and Kumar [9] for details of the causal history approach and the MIWTR technique.

3.2.5 Classifying causal structure evolution by clustering DAGs

To determine the different types of causal structure embedded in the dynamics of a set of DAGs for time series, we generate clusters of DAGs to find patterns of similar behavior. In this study, each DAG shows the near-instantaneous (3-min average) state of the dynamics of the multivariate system. To compare a pair of DAGs, we use a distance metric proposed by Malmi, Tatti, and Gionis [115], as briefly summarized. For a pair of DAGs, G_l and G_m, the Kendall-tau distance $K(G_l, G_m)$ between the DAGs is the number of all concordant and discordant vertex pairs when comparing them. A metric to compare DAGs based on the configuration of vertices is possible given that all DAGs have the same vertices (nodes), and therefore, a vertex (link) correspondence between any pair of DAGs is known. For a pair of DAGs,

$G_l = (V, E_l)$ and $G_m = (V, E_m)$, two vertices of V are denoted as i and j, and $i < j$ if $i \in V$ precedes $j \in V$. As mentioned before, V is fixed and common between the two DAGs being compared. Two edges connecting the previous vertices are denoted as $e = (i, j)$ and $f = (j, i)$, where f is e with reversed direction. Then, the distance between a pair of DAGs is denoted by $K_{i,j}^{(p,q)}(G_l, G_m)$, and is the sum of penalties associated with the vertices i and j, with respect to the fixed parameters p and q, where $0 \leq p, q \leq 1$, as follows [115]:

$$K^{(p,q)}(G_l, G_m) = \sum_{\substack{(i,j) \in V \times V \\ i<j}} K_{i,j}^{(p,q)}(G_l, G_m) \tag{3.3}$$

Malmi, Tatti, and Gionis [115] consider four cases for penalization as follows:

1. $(e \in E_l$ and $e \in E_m)$ or $(f \in E_l$ and $f \in E_m)$, meaning that if both DAGs, G_l and G_m, agree on e, then the distance between the DAGs is zero: $K_{ij}(G_l, G_m) = 0$.

2. $(e \in E_l$ and $f \in E_m)$ or $(e \in E_m$ and $f \in E_l)$., that is if the DAGs G_l and G_m completely disagree on e, then the distance is equal to 1, $K_{ij}(G_l, G_m) = 1$. This case cannot occur for DAGs from time series, as shown with a toy example in appendix A.5.

3. $(e$ or $f \in E_l$ and $e, f \notin E_m)$ or $(e$ or $f \in E_m$ and $e, f \notin E_l)$., that is an edge between i and j exists in one graph but not in the other. In this case we set $K_{ij}(G_l, G_m) = p$

4. $e, f \notin E_l$ and $e, f \notin E_m$., in which there is no edge between i and j in either of the graphs. We set $K_{ij}(G_l, G_m) = q$. Therefore, $K(G, G) = 0$, however this property does not hold for $q > 0$ as penalty q is induced by edges that are missing from both input graphs. The motivation for setting $q > 0$ is that otherwise two empty DAGs are as similar as two identical complete DAGs even though one might argue that the latter two have much more in common than the empty DAGs.

In appendix A.5, we provide an example for the estimation of the distance metric K when implementing the penalization's for DAGs for time series. From now on, we omit p and q from the notation by specifying $p = 1$ and $q = 0$, and write $K(G_l, G_m)$ instead of $K^{(p,q)}(G_l, G_m)$. We do not penalize using q, given that the DAGs obtained from our data are sparse DAGs and we do not expect to have pairs of identical empty (i.e., fully-disconnected) or complete (i.e., fully-connected) DAGs. Note that even though the definition of $K(G_l, G_m)$ is based on an ordering of vertices, the final result does not depend on that. To test the performance of the $K(G_l, G_m)$ metric, we evaluated two scenarios that are included in appendix A.6.

In this study, we use the distance measure $K(G_l, G_m)$ between DAGs for clustering. To create clusters of DAGs with similar causal structure, we use a method of DAG aggregation to determine a cluster's. Given a set of DAGs, $G_l, \ldots, G_M$, and a candidate centroid $C = (V, E)$, we can estimate the cost of not having the edge (i, j) in the centroid, as $b(i, j)$ [115]:

$$b(i,j) = \sum_{m=1}^{M} K_{ij}\left(G_m, G_{\text{empty}}\right) \tag{3.4}$$

where G_{empty} is a DAG with no edges. Similarly, $w(i,j)$ is the cost of having (i,j) as an edge in the centroid,

$$w(i,j) = \sum_{m=1}^{M} K_{ij}\left(G_m, G_{\text{full}}\right) \tag{3.5}$$

where G_{full} is a fully connected DAG without any contemporaneous links since they do not reflect causal relation. The estimation of G_{full} is included in appendix A.7. This gives us [115]:

$$\sum_{m=1}^{M} K\left(G_m, C\right) = \sum_{e \in E} w(e) + \sum_{e \notin E} b(e) \tag{3.6}$$

Now, consider $P = \{e \in V \times V \mid w(e) < b(e)\}$ as the set of edges that ideally has the centroid. We assume that the centroid edges are all in $P, E \subseteq P$ [115]. However, since P may contain cycles, we cannot have all of those edges. For each edge e not included in the centroid, we account for *regret* [115], $r(e) = b(e) - w(e)$, to be the difference to pay for not having the edge in C. We can write the *regret* cost as [115]:

$$\sum_{m=1}^{M} K\left(G_m, C\right) = \sum_{e \in P} w(e) + \sum_{e \notin P} b(e) + \sum_{e \in P \setminus E} r(e) \tag{3.7a}$$

$$\sum_{e \in P \setminus E} r(e) = \sum_{e \notin P} b(e) - \sum_{e \in P} w(e) \tag{3.7b}$$

The first two terms in equation 3.7**a** do not depend on C. Therefore, to find the optimal centroid we need to minimize the third sum in equation 3.7**a** [115]. This is used in a Greedy algorithm [115] to estimates the optimal centroid of a set of DAGs. Briefly explained, in the Greedy algorithm we order the edges based on *regret*, where edges with the smallest *regret* go first [115]. Note that we only need to consider edges that have appeared in at least one of the input DAGs, since $w(e) < b(e)$ can not hold for the other edges. Then we keep adding edges to the DAG denoted as centroid in order, ignoring the edges that create cycles. In Malmi, Tatti, and Gionis [115] Greedy algorithm's approach outperformed another algorithm for which there is a constant approximation guarantee. Therefore, in this study, we use the Greedy algorithm to estimate the cluster's centroid. For more details in the Greedy approximation of DAG aggregation, we refer the reader to Malmi, Tatti, and Gionis [115].

Our purpose for clustering DAGs is to find groups of similar causal structures, which we consider as a representative of the near-instantaneous behavior of the system. For clustering DAGs, we use a k-means approach, which helps organize the DAGs into clusters, without imposing a hierarchical structure [124]. To update the cluster centroid in each

iteration we use the previously described Greedy algorithm. Given a set of M DAGs G_1,...,G_M, we find k number of clusters denoted by P_1, P_2..., P_k, with centroids denoted by C_1, C_2..., C_k, such that:

$$\sum_{i=1}^{k} \sum_{G \in P_i} K^{(p,q)}(G, C_i) \tag{3.8}$$

is minimized [115]. Therefore, from a set of centroids, we group the input DAGs into clusters minimizing the distance among them. After the clusters are formed, we find the centroid of each cluster. The initialization of the algorithm is random and updated after several restarts using the Greedy algorithm, which updates the centroids [115]. We run five restarts with random initial cluster assignments and select the run which minimizes the clustering cost. A challenge using a k-means type of approach for clustering is the random selection of the initial centroids, which have a significant impact on the results [125]. We select a seed to initialize the Mersenne Twister random number generator [126], resulting in the same random numbers after each call to the random number generator, guaranteeing the reproducibility of the results. We use the Nominal-k method [127] as seed, a rule of thumb based on the number of elements to be clustered, n, as follows:

$$s = \sqrt{n/2} \tag{3.9}$$

While the selection of the seed for the Mersenne Twister algorithm used to shuffle is arbitrary, the seed only determines where we start in a very long list of pseudorandom numbers uncorrelated with each other, that is used as a generator.

3.2.6 Determining an unbiased number of clusters

One of the challenges of using k-means approach for clustering is the selection of the number of clusters, k, which can be subjective. In Malmi, Tatti, and Gionis [115] k is an input parameter, which is set along with p and q. However, we require unbiased criteria to determine k to explore causal inference from the system's connectivity. We achieve this by proposing a Dunn index for DAGs, which is based on Dunn's [128] work and it is calculated as the ratio of the minimum of inter-cluster DAG distances and the maximum of intra-cluster DAG distance using the $K(G_l, G_m)$ metric. This proposed Dunn index for DAGs measures the compactness of the generated clusters by accounting for the maximum intra-cluster distance, which is the maximum distance in between DAGs of a cluster. Also, the Dunn index for DAGs considers cluster separation using the inter-cluster distance, that is the minimum distance between clusters of DAGs. Let $\delta(C_i, C_j)$ be the inter-cluster distance, between clusters C_i and C_j. Therefore, the Dunn index (DI) for DAGs for m cluster is defined as follows:

$$DI_m = \frac{\min\limits_{1 \leqslant i < j \leqslant m} \delta(C_i, C_j)}{\max\limits_{1 \leqslant l \leqslant m} \Delta_l} \tag{3.10}$$

where Δ_l refers to the intra-cluster distance for each cluster l, which is estimated by:

$$\Delta_l = \sigma_l^2 = \frac{\sum l^2(\bar{D}, D_l)}{n_l} \quad . \tag{3.11}$$

Here, the distance between the cluster's centroid and each DAG is denoted by D_l, and n_l refers to the number of elements in cluster k. For the Dunn index, a higher value is associated with better clustering. It is then used to determine the optimal number of clusters k by looking at the number of clusters: the k for which the Dunn index is the highest is then chosen as the number of clusters. The Dunn index helps to identify sets of clusters that are compact, with a small variance between members of the cluster, and well separated, where the means of different clusters are sufficiently far apart, compared to the within-cluster variance.

3.3 Results

3.3.1 Causal structure and their evolution

We use a series of consecutive DAGs to explore the evolution of the causal structure that reflect the temporal interdependencies in a multivariate system. We subdivide the multivariate time series of high frequency instantaneous data measured at 10Hz from 10:46:53 CST to 14:01:53 CST (Central Standard Time) in non-overlapping contiguous 3-minutes windows, resulting in 66 consecutive windows of 1800 data points per variable for each window on August 18 and 21, 2017. We estimate the DAGs for time series using Tigramite algorithm [43], [51]–[53], which is a modified PC algorithm (i.e., Peter and Clark algorithm)[10]) for constructing the DAG representation of time series by using a MIT-based conditional independence test to remove the spurious relationship between two nodes [51]. Each conditional independence test is performed based on 100 samples with a significance level = 0.05. For the empirical estimation of all information-theoretic measures, Tigramite uses the k-nearest-neighbor (knn) method [121], [129]. In this study, knn is used to compute CMI with the nearest-neighbor parameter $k = 100$. A high value of k reduces the variance of the estimated CMI [121], however, it is recommended that k be 5% - 50% of the time series length [51]. The maximum Markovian time lag used for establishing the directed edges between lagged variables is set to 9. When each DAG is plotted (Figures 3.2e and 3.2f), the thickness of each edge represents the coupling strength between two connected nodes, which is given by MIT [43].

After we estimate the DAGs, we use their sparse adjacency matrices to analyze the evolution of the network structure.

To obtain each adjacency matrix, we use the MIT weighted values between each pair of lagged nodes, which are connected when their MIT is larger than a significance threshold that is estimated using the shuffle test as described in Runge, Heitzig, Petoukhov, *et al.* [43]. We developed a modified Tigramite plotting script to extract the adjacency matrices from the previously constructed DAGs using the NetworkX Python package [130]. After creating one group of adjacency matrices, we estimate the DAGs clusters using the k-means type of algorithm described in subsection 3.2.5. While clustering, each DAG is treated as an element inside a cluster of similar DAGs. Using the distance metric between DAGs, $K(G_l, G_m)$, we estimate different numbers of clusters, k, showing variability in the composition of the DAGs (Figure 3.3, red and blue dots, respectively).

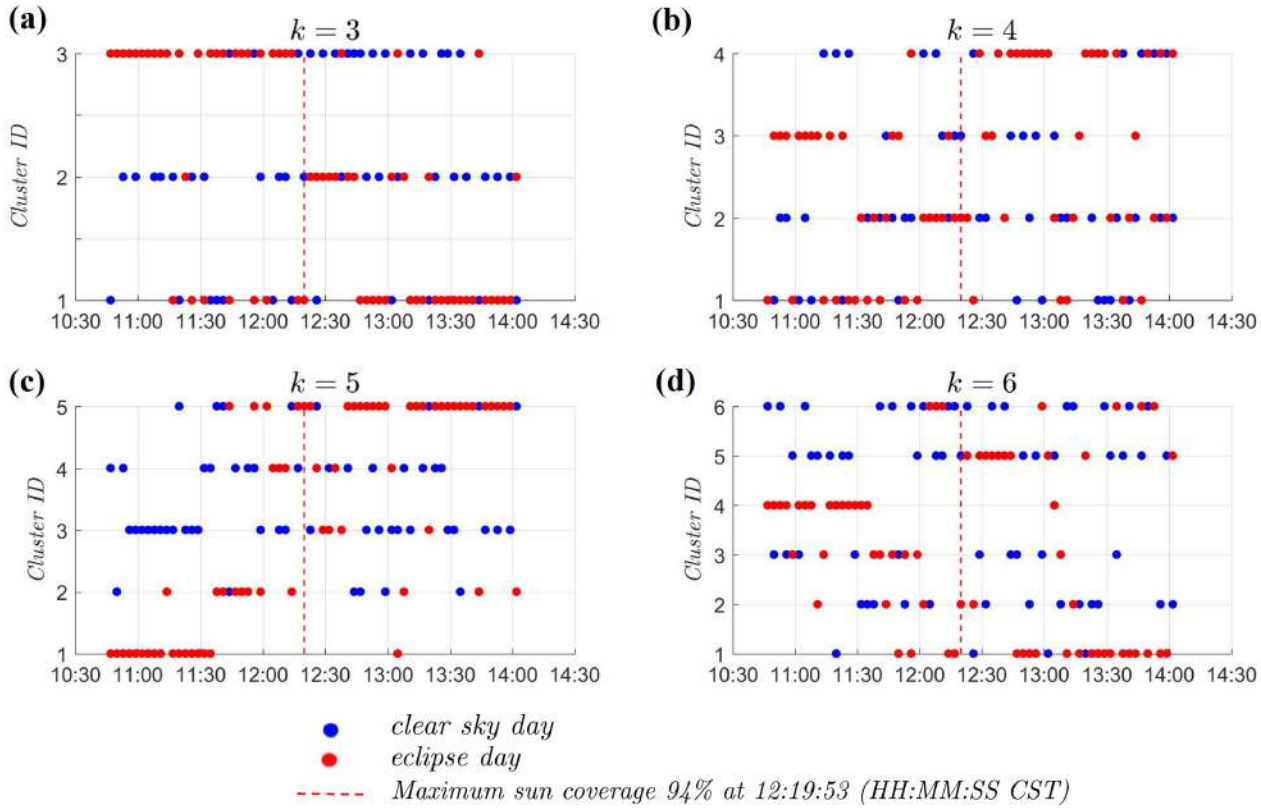

Figure 3.3: Time-ordered distribution of clustered DAGs, for a clear sky (blue) and the eclipse (red) day, for different number of clusters, k: (**a**) k =3, (**b**) k =4, (**c**) k =5, and (**d**) k =6.

We observe that each cluster is a mix of the DAGs from both days. However, when the number of clusters increases, for instance for $k \geq 5$, we observe that a cluster is formed uniquely by DAGs of the early dynamics of the partial eclipse (Figure 3.3c and d, cluster #1 and #4, respectively). However, since the selection of the optimum number of clusters, k, can be subjective, we use the Dunn index for DAGs as a quantitative criterion to select the optimal number of clusters (Subsection 3.2.6). The results from the Dunn Index (Figure 3.4c) indicates that $k = 3$ offers the most parsimonious classification of the structure of the DAGs.

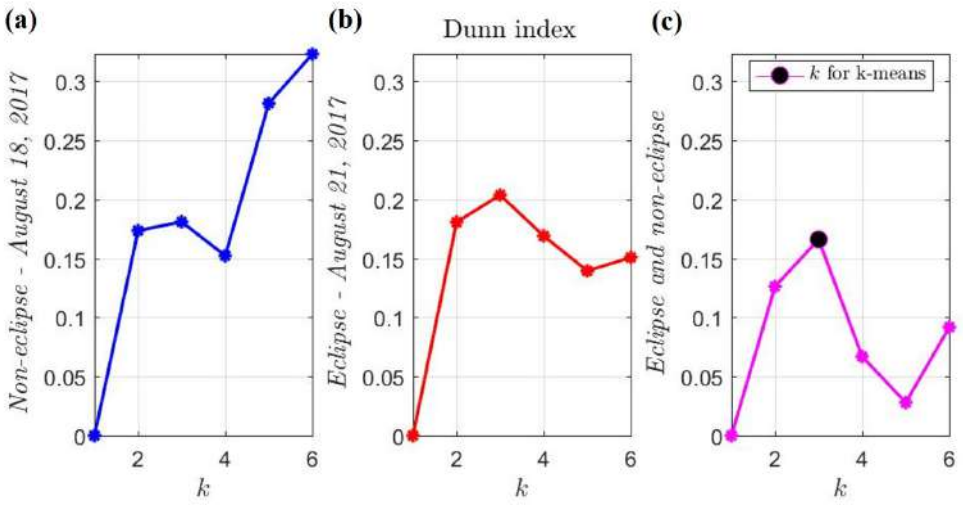

Figure 3.4: We use the Dunn index for DAGs to determine the number of clusters, k, for (**a**) a clear sky day, August 18, 2017, (**b**) the eclipse day, August 21, 2017, and (**c**) the combined DAGs from both days. The selected number of clusters $k = 3$ (black filled dot) refers to the maximum Dunn Index when using all DAGs, which is also the maximum and the first local maximum in (**b**) and (**a**), respectively.

Does the causal structure change in time and are those patterns of connectivity statistically distinguishable? Using the clusters formed when $k = 3$ we plot the time-ordered distribution of the DAGs in clusters for a couple of metrics and indexes (Figure 3.5), as described below. We observe that clusters of DAGs with similar dynamics emerged at turbulence scales, creating patterns of similar behavior (Figure 3.5, cluster #1 and #3, respectively).

First, we estimate the distance $K\,(G_l, G_m)$ between each pair of consecutive DAGs in the sequence, i.e., G_{t-1} and G_t for all t (Figure 3.5a). Note that G_{t-1} and G_t refers to a DAG at a lagged time $t-1$ with respect to the current DAG at time t. In this case, a larger distance between the pair indicates rapid change in structure. On the other hand, if the distance decreases between two consecutive DAGs, they become more similar to each other. If there is no change in the distance, the consecutive DAGs are identical in terms of their causal structure. We observe that, in general, the distance between consecutive DAGs is smaller during the eclipse (Figure 3.5a), meaning that the DAGs are more similar to one another in comparison to the DAGs in the clear sky day. Next, we account for the distance between each DAG and a hypothetical fully connected DAG (Figure 3.5b). In this case, a larger distance between them is an indicator of a sparser DAG. On the other hand, the smaller the distance, the denser the DAG is, and therefore, the more similar to the fully connected DAG. We observed that, in general, the DAGs of the clear sky day have a smaller distance to the fully connected DAG than the DAGs in the eclipse day. Therefore, the DAGs of the clear sky day are, in general, denser and therefore, more similar to a fully connected DAG. During the clear sky day, the DAGs show, in general, more interdependencies among variables than during the eclipse day, meaning that the system during the clear sky day exhibits a more dynamic behavior and is less sparse than that on the eclipse day.

To facilitate the recognition of the patterns, we propose a couple of indices to characterize the evolution of the causal

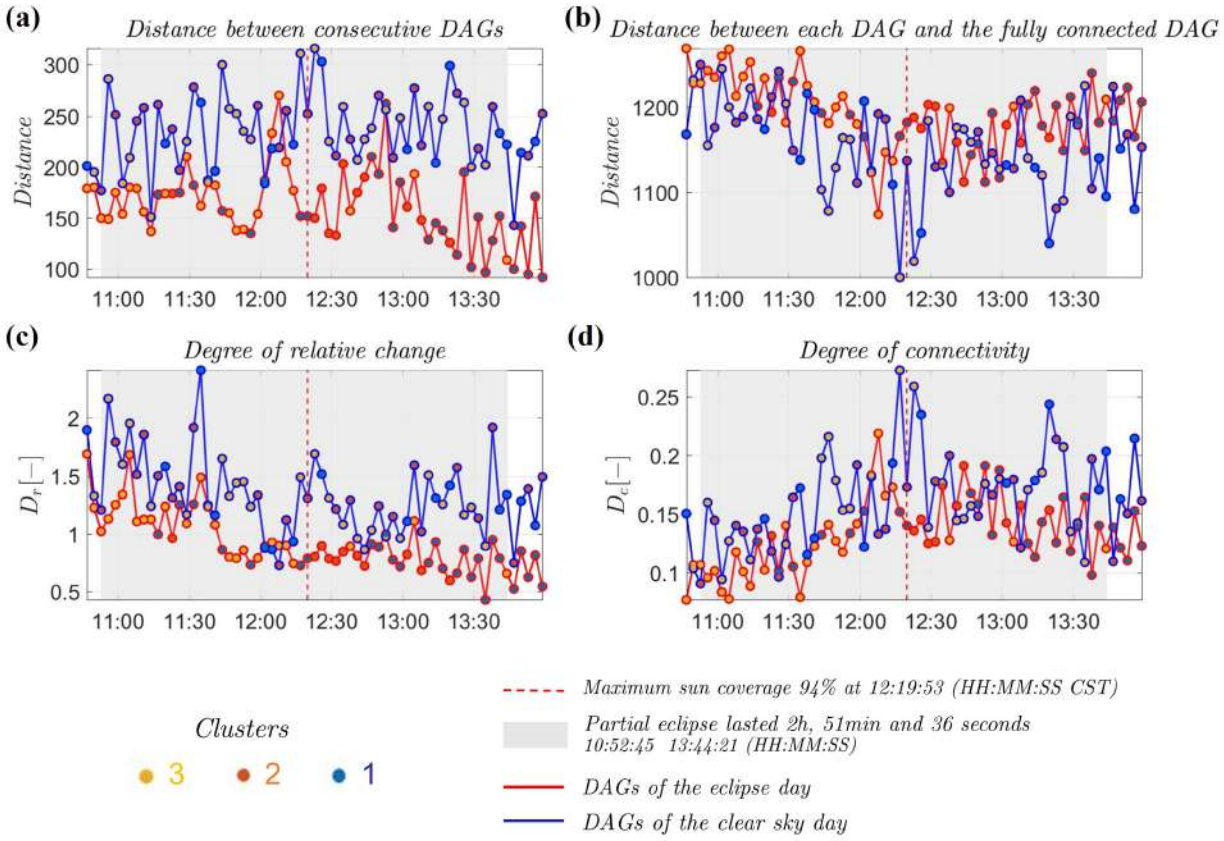

Figure 3.5: Metrics and indices show the evolution of the causal structure of the system for the time-ordered clustered DAGs. The filled dots refer to DAGs in cluster #1 (yellow color), cluster #2 (red color), and cluster #3 (blue color). The lines connect DAGs of the same day, for the clear sky (blue) and eclipse day (red). The metric are the (**a**) distance between two consecutive DAGs, and the (**b**) distance between each DAG and a hypothetical fully connected DAG. The index refer to the (**c**) degree of relative change, D_r (Equation #3.12), and the (**d**) degree of connectivity, D_c (Equation #3.13).

structure of DAGs. We consider two types of evolutionary indices: the degree of relative change, D_r, and the degree of connectivity, D_c. For a sequence of consecutive DAGs, the degree of relative change, D_r, is estimated by the ratio between the distance of any two consecutive DAGs, G_{t-1} and G_t, and the number of edges, e, in the lagged DAG, G_{t-1}, as follows:

$$D_r = \frac{K(G_t, G_{t-1})}{e_{G_{t-1}}} \tag{3.12}$$

Note that D_r is always positive. It is less than 1 when the distance of any two consecutive DAGs is larger than the number of edges, e, in the lagged DAG, G_{t-1}, and more than 1 otherwise. During the eclipse day (Figure 3.5c), D_r is less than 1 near the time of the maximum obscurity and after that, which coincides with the formation of patterns of similar consecutive DAGs that do not change much in terms of their dynamics (Figure 3.5c). On the other hand, during the clear sky day, D_r is usually larger than 1. Therefore, the interdependencies shown by the DAGS of the clear sky day reflect a more dynamic behavior of the system than during the eclipse day.

In the case of the degree of connectivity, D_c, it accounts for the ratio of the distance, $K(G_t, G_{\text{full}})$, estimated between each sparse DAG from time series, G_t, and a hypothetical fully connected DAG, G_{full}, and the number of edges in the fully connected DAG, $e_{G_{\text{full}}}$, as follows:

$$D_c = 1 - \frac{K(G_t, G_{\text{full}})}{e_{G_{\text{full}}}} \tag{3.13}$$

The estimation of the number of edges in the fully connected DAG is described in the appendix A.7. Note that D_c is always positive because the number of edges in G_{full} is always larger than the distance between any pair of DAGs formed by a sparse DAG and a fully connected DAG. Given that G_{full} is a constant, a larger D_c refers to a denser DAG. In other words, the closer D_c is to 1, the better connected the DAG. We observe that the DAGs from time series are sparse, with D_c values ranging from 0.05 and 0.27, approximately. Furthermore, we observe that, in general, the connectivity of the DAGs is lower for the eclipse day when compared to the behavior of a clear sky day, where a larger D_c is observed (Figure 3.5d). Therefore, we find that the perturbation created by the solar eclipse causes a slight decrease in the connectivity of the dynamics of turbulence with respect to the dynamics during a clear sky day, where D_c is generally higher. Moreover, we observe that the degree of connectivity fluctuates within a range of variation, that can be slightly altered and recovered, even under the effect of an intervention on the whole system dynamics such as the solar eclipse.

For the selected number of clusters, $k = 3$, each cluster's centroid can be seen as representing the average dynamics within a cluster, which are estimated as described in subsection 3.2.5. However, such centroids are synthetic representations that are not associated with a time series of observations. For this reason, we look for the DAG of the time series that is closest to the centroid of each cluster. Moreover, the closest DAGs to the centroid of the clear sky and eclipse day (Figure 3.6, blue and red dots above the closest DAG to the centroid, respectively), are useful to determine the dynamics associated with the same cluster. Some measures of statistical dispersion for each cluster are also useful

to characterize the configuration of the clusters (Figure 3.6). For example, the average distance of all DAGs within a cluster to the centroid of that cluster is approximately the same (i.e., ≈ 125), although, the number of DAGs inside each cluster varies. Cluster #3 has the largest number of DAGs within the cluster, followed by cluster #1 and #2. In addition, cluster #2 has the closest DAG to the synthetic centroid across clusters. To plot the synthetic centroids (Figure 3.7), we modified a Tigramite [43], [51]–[53] plotting scrip where we use the resulting adjacency matrix of the centroid of each cluster to generate the synthetic DAGs. Note that the thickness of the edges in the synthetic DAGs associated with the centroids does not represent a differential weight on the connectivity between a pair of nodes, but merely the causal connectivity between them. The synthetic centroids (Figure 3.7a, b, and c) show the main characteristics of the causal dependence among variables for each of the three clusters.

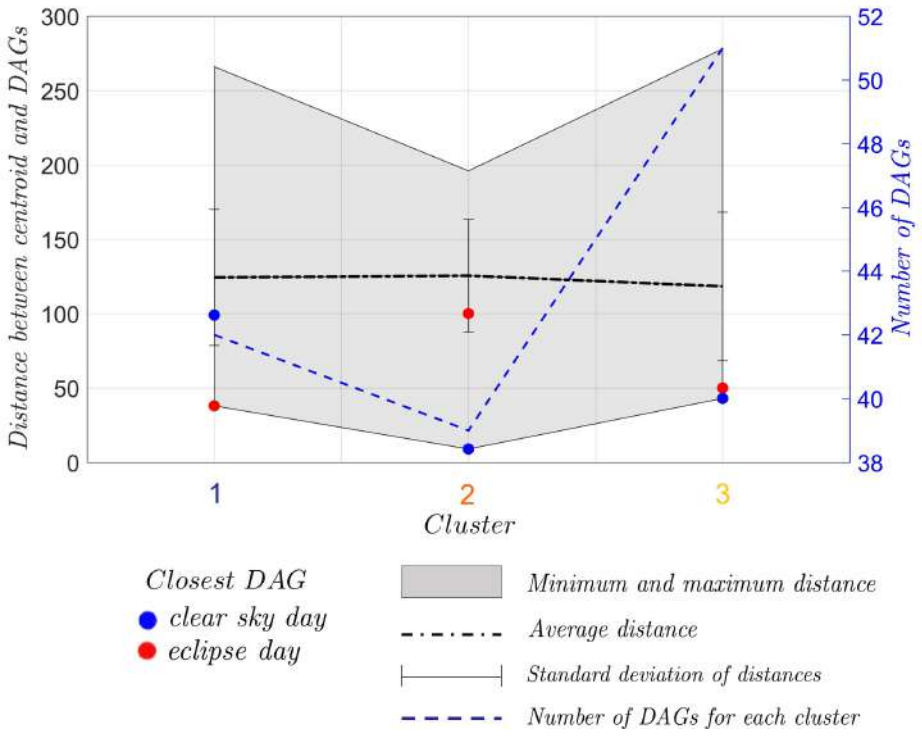

Figure 3.6: Distance between the centroid of a cluster and the DAGs inside the cluster. For clusters #1 to #3, the closest DAG to the centroid (i.e, minimum distance) can either be from a clear sky (blue) or the eclipse (red) day. The dot above refers to the closest DAG from the opposite day inside the cluster. The minimum and maximum distance between the centroid of a cluster a the closest and farthest DAG (i.e., shadow gray area), along with the average distance (i.e., black dash-dotted line), the standard deviation (i.e., error bars), and the number of DAGs inside each cluster (i.e., blue dashed line), provide insights of the composition of each cluster which are formed from a combination of DAGs from both days.

3.3.2 Interpreting changes in the causal structure among clusters

For the physical interpretation of the dynamics of each cluster, we note that the variables in the DAGs can be classified in three groups based on both their nature and role on the dynamics of the system. First, the horizontal and

vertical wind components, WS and U_z, reflect the dynamic character of the atmosphere. Second, the temperature, T, reflects the thermodynamics. Third, CO_2 and H_2O reflect the scalar transport at the biosphere-atmosphere interface. The behavior of the interactions among these three different groups facilitates the physical interpretation of the causal interactions, that is, how are the three clusters different in terms of their dynamics? We characterize the differences among the clusters using the DAGs from the time series that is closest to the centroid that belongs to each cluster. However, to physically interpret the causal dynamics of each cluster we need to identify the distinguishable features of each cluster. In the DAGs from time series, the strength of the connectivity between variables is shown by the thickness of the edges linking the source and target variables. Therefore, in a DAG, thicker edges refer to stronger dependencies between the lagged stages of the variables ([43], [118]). For the high frequency data at 10Hz, the x-axis in a DAG shows a lagged time with respect to a current time t expressed by $t - \tau$, where τ is a positive time lag equal to 0.1s. Therefore, each illustrated DAG shows the connectivity of the multivariate set for up to one second (Figure 3.7).

For cluster #3 (Figures 3.7d and 3.7g, for the eclipse and the clear sky day, respectively), the dynamics of the horizontal WS, and vertical wind speed, U_z, are largely self-dependent. However, the thermodynamics, T, has a weak causal influence on WS. However, if T strongly influences the scalar transport of CO_2 and H_2O. That is, thermodynamics is the primary causal influence on the vertical scalar transport. Furthermore, CO_2 and H_2O also show causal dependence on each other at one time step lag. This finding is consistent with the strong coupling of photosynthesis and transpiration during the period of high primary productivity. However, it is intriguing that this dependency is manifest in the data at high frequency in spite of the mixing through the large flux footprint associated with measurements at 25m height [26].

For cluster #1 (Figures 3.7f and 3.7i, for the eclipse and the clear sky day, respectively), we observe that the dynamics are similar to those as in cluster #3. Similarly, the dynamics of the horizontal WS, and vertical, U_z, wind speed are largely self-dependent. However, we observe that the vertical wind U_z has some influence on temperature, T, and in the horizontal wind, WS, while the thermodynamics, T, are strongly coupled to the scalar transport of H_2O and CO_2.

For cluster #2 (Figures 3.7e and 3.7h), the dynamics of the horizontal WS, and vertical wind speed, U_z, are decoupled from the thermodynamics, T. Therefore, the scalar transport of CO_2 and H_2O depends only on the thermodynamics with a weak coupling between them, which can be attributed to photosynthesis and transpiration. An additional distinguishable feature in cluster #2 is the weak self-dependency of temperature T, which can lead to a complete self-decouple (Figure 3.7e). However, even if the temperature shows a self-decoupled behavior during the eclipse, it still sustains the dynamics of the scalar transport of H_2O and CO_2. Therefore, cluster #2 is mainly characterized by a decouple between the two main attributes controlling the dynamics of the whole system at turbulence scales: the dynamic influence of the atmosphere given by the wind speed, and the influence of the thermodynamics, which is strongly associated to the scalar transport.

In this causal inference analysis of the dynamics of the system at high frequency, we distinguished the predominant

role of the influence of the thermodynamics on the scalar transport, for both H_2O and CO_2, from the influence of the dynamics of the wind. While the general concept is that the vertical wind transports the scalars in the biosphere-atmosphere interface, we here observe that the thermodynamics have a more relevant role for the given conditions. At 25m height, the roughness effect from the surface is not a major control in the turbulent transport of scalars, given that our results suggest that the influence of the thermodynamics on the buoyant movement of scalars is more dominant at this high resolution. In addition, the influence of the wind on temperature was not observed, while the thermodynamics are more likely to occasionally weakly influence the horizontal and vertical wind speed (Figure 3.7, cluster #3 and cluster#1, respectively).

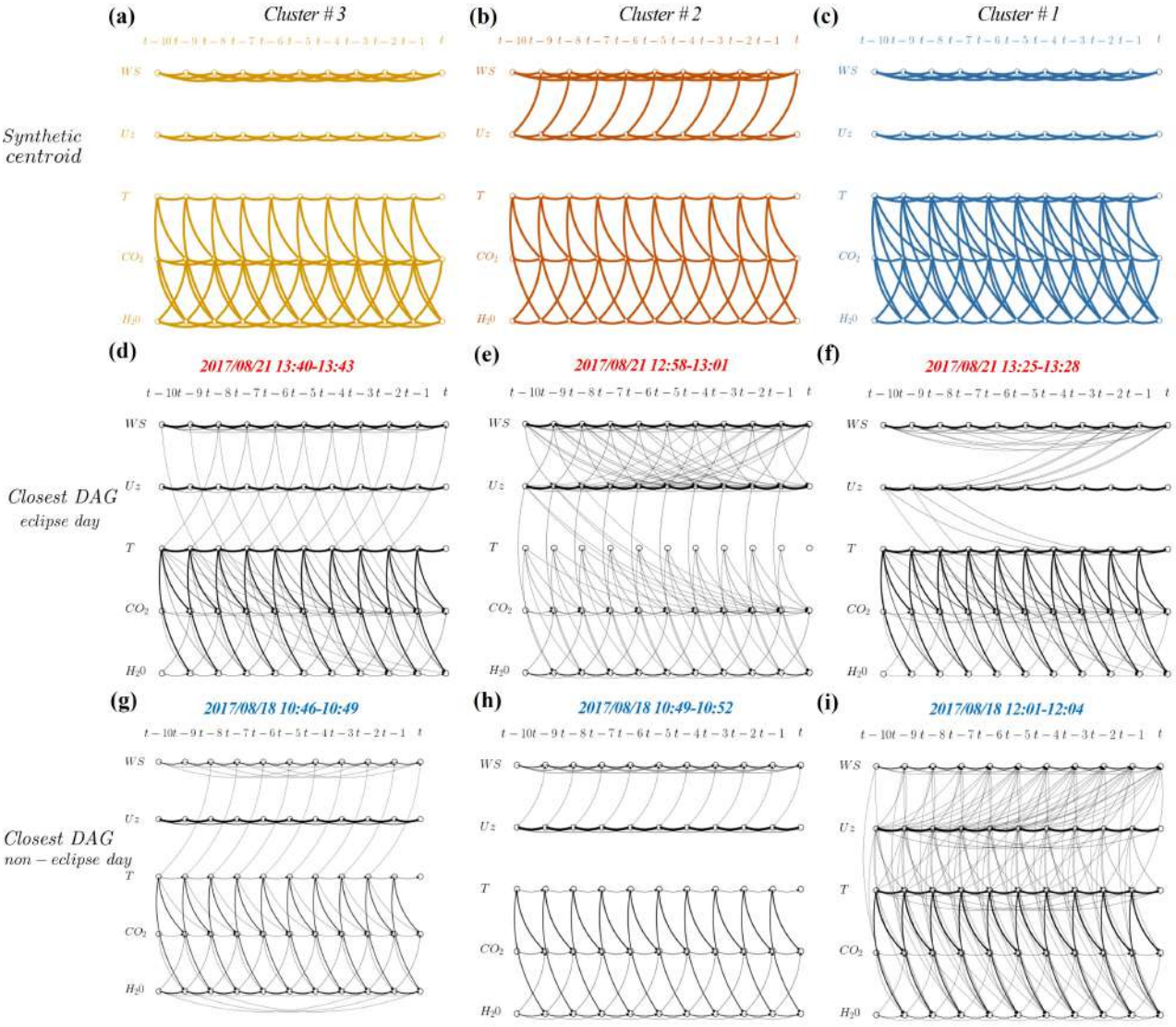

Figure 3.7: Synthetic centroids and closest DAGs from time series to each cluster's centroid. The synthetic centroid shows the average of the dynamics of (**a**) cluster #3 (in yellow), (**b**) cluster #2 (in red), and (**c**) cluster #1, in blue. The closest DAGs from time series to each cluster's centroid for the eclipse (**d,e,f**) and the clear sky day (**g,h,i**).

Causal dependencies in a multivariate system can be considered as a subset of all dependencies that can be found when analysing time-series from observations. Specifically, causal dependencies are carried out by the propagation of fluctuations and are characterized by being temporally asymmetric, meaning that one variable influences the future of another variable asymmetrically in time [54], [55].

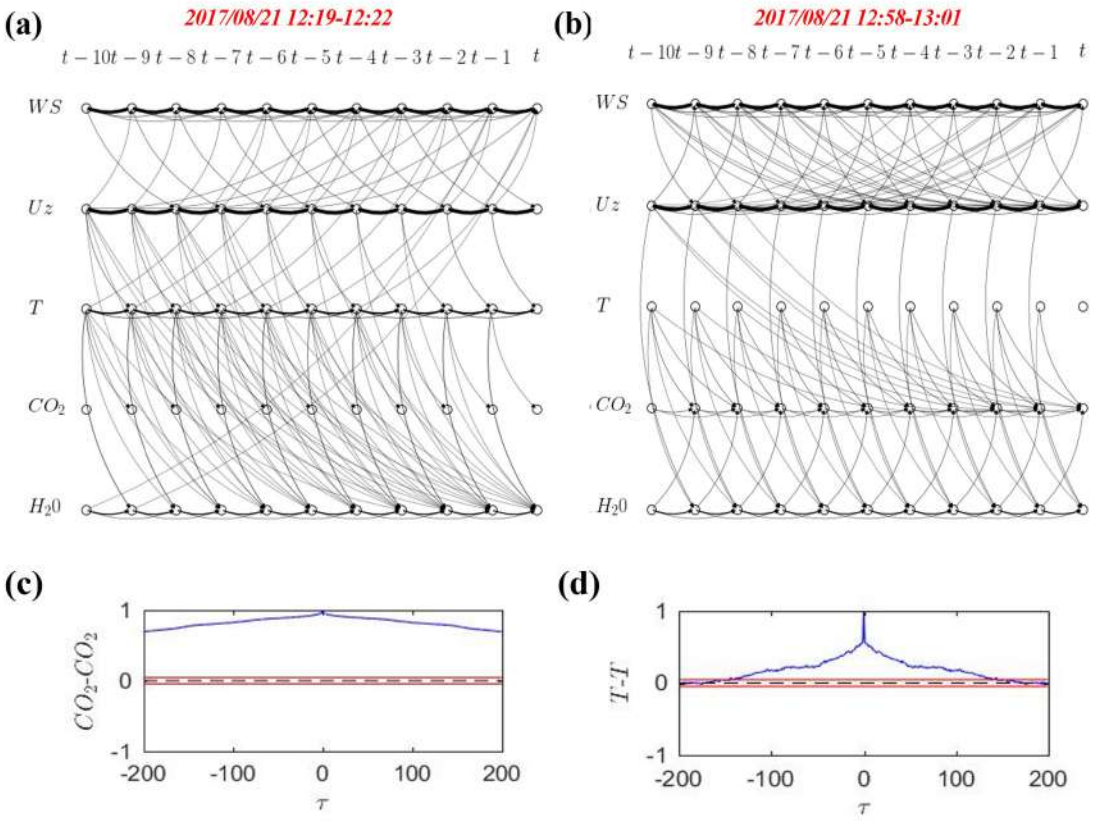

Figure 3.8: DAGs that show the apparent absence of self-feedback on the dynamics of variables at turbulence scales. A unique case for (**a**) CO_2 at 12:19-12:22 on August 21, 2017, and an example of the more frequent case of absence of self-feedback for (**b**) temperature, as observed at 12:58-13:01 on August 21, 2017.

For the cases where a variable seems to be self-decoupled, such as in the case CO_2 and T (Figure 3.8a and 3.8b, respectively), the causal dependencies might be near-instantaneous, meaning that they occur before a 0.1s time lag, or that the influence might come from a longer time lag than $\tau = 9$, causing the links to disappear from the causal graph. However, a lagged influence out of $\tau = 9$ seems to be unlikely given the rapid nature of the interactions in the system. To further explore this self-decoupling behavior, we can alternatively use the time series to explore other dependencies using a cross-covariance analysis to verify the nature of the behavior observed in our causal analysis for the time series of the centroids of the clusters (Figures A.10 to A.15, in appendix A.8). For CO_2 (Figure A.16), the correspondence behavior of the auto-covariance of CO_2 shows a strong instantaneous auto-correlation at time zero that very slowly dies out as the time lag increases. In the case of the self-decoupled T (Figure A.13), we observe a very

strong auto-correlation at time zero that rapidly dies out after a few seconds. Therefore, this results suggest that the apparent self-decoupling behavior shown in the DAGs for CO_2 and T is due to the near-instantaneous nature of the self-dependence, which is not detectable using observations at 10Hz.

By exploring the cross-covariance across clusters (Figures A.10 to A.15, in appendix A.8), we find correspondence to the characteristic causal behavior observed between CO_2 and H_2O. Also, we observed that the horizontal wind speed WS, is mostly uncorrelated to T, CO_2 and H_2O, which agrees with our causal analysis.

In addition, we observed a particular behavior in the DAG with the apparent self-decoupled behavior of CO_2. In general, during day-time we expect the influence of the thermodynamics on the scalar transport of CO_2 and H_2O, as well as from the dynamics of the wind, both from the horizontal, WS, and vertical wind speed U_z. However, we would not expect the scalars to influence and modify the dynamics, unless they have the ability to modify material properties. That is precisely what we observed in Figure 3.8a. During this window in the eclipse day, our results suggest that the presence of water in a nearly stagnant environment due to the effect of the solar eclipse, might have changed the density in a significant way, resulting in the temporal lagged influence of H_2O on the horizontal wind speed, WS, which we did not expect to observe at this fine resolution.

3.3.3 Inferring the creation of patterns of similar behavior

After identifying the distinguishing attributes of each cluster (Subsection 3.3.2), we can now causal infer the behavior of the evolution of the dynamics of the system, as observed in the time-ordered distribution of the DAGs for the three clusters (Figure 3.9). On a clear sky day (Figure 3.9a), most of the dynamics first exhibit the behavior of cluster #2, characterized by the decoupling of the wind from the scalar fluxes, H_2O and CO_2, causally driven by thermodynamics. Over time, the dynamics of the system tend to migrate towards the characteristic behavior of cluster #3, characterized by a high connectivity of the system, with a weak coupling of the wind to the dynamics of the scalar transport of H_2O and CO_2, and a stronger control of the thermodynamics. Therefore, we can infer that during a clear sky day the dynamics of the system tends to easily fluctuate between the states of decoupling and weak coupling, with some tendency to move from the behavior of cluster #2 to cluster #3 (Figure 3.9a).

On the other hand, during the eclipse day (Figure 3.9b), many of the DAGs that correspond to the behavior of the cluster #3 occur before the maximum obscurity, characterized by a high connectivity of the system, where the scalar transport of H_2O and CO_2 is mainly led by thermodynamics but also by the weak influence of the wind. As the solar eclipse reaches the maximum obscurity and right after, the system transitions to a cluster #2 behavior, characterized by the decoupling of the dynamics of the wind from the thermodynamics and the scalar transport of H_2O and CO_2. As the partial solar eclipse progresses the system reaches a cluster #1 behavior, where the main characteristic is the recovery of the influence of vertical wind speed on H_2O and CO_2. Based on these results, we infer that the intervention

on the dynamics of the system created by the eclipse forced the system to behave differently than during a clear sky day, creating more consistent patterns of similar behavior over time (Figure 3.9b).

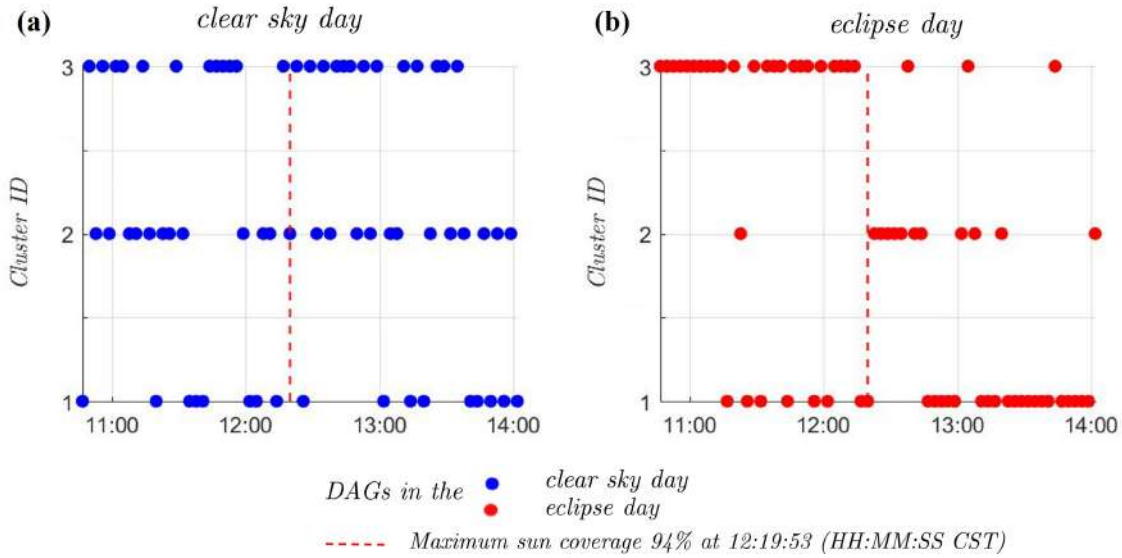

Figure 3.9: Causal inference to determine the tendency of the behavior of the system over time. The time-ordered distribution of the clustered DAGs for $k = 3$, show different patterns of behavior for the (**a**) clear sky (blue dots) and the (**b**) eclipse day (red solid dots).

3.4 Discussion

3.4.1 Considerations in the analysis of the causal structure

In this study, we consider three aspects which can help clarify the interpretation of our results. First, when comparing DAGs to explore the evolution of causal structures, the change in its structure is a result of changes in dependency among the variables. Second, these interdependencies are the result of the propagation of fluctuations or variations among variables. Third, the causal dependencies are generally asymmetric, meaning that the strength of influence of a variable such as $T(t-1) \rightarrow H_2O(t)$ is generally not the same as $H_2O(t-1) \rightarrow T(t)$.

An additional aspect refers to the pertinence of the use of data from a solar eclipse when studying the evolution of the causal structure of turbulence at the biosphere - atmosphere interface. We use such a unique intervention as part of our analysis, not to emphasise on the behavior of ecohydrological processes for this particular condition, but because for such a short period the dynamics of the whole system are subjected to the rapid decline and recovery of solar radiation.

Turbulence at the biosphere-atmosphere interface creates a nonlinear system of multivariate interdependencies with a changing degree of coupling among its components. The rapid interactions among variables at high frequency, such as wind speed, air temperature, water vapor, and CO_2, create a network of complex interactions where information

flows among variables at lagged times defines its causal structure. The analysis indicates that over time, these networks of causal dependencies propagate fluctuations that produce patterns of similar behavior at high frequency under the influence of a strong intervention of the dynamics of the system.

Natural systems that evolve with and learn from interactions exhibit both structural order and dynamical chaos [131]. Order, and therefore structure, is the foundation of communication between variables. Similar to how an organism's form is a functional manifestation of its ancestor's evolutionary and its own developmental memories [131], the causal structure of a complex system is a functional manifestation that encodes the evolution of its causal history.

Our results show the ability of our approach to determine the patterns of the inherent causal structure in turbulence. The centroids for each cluster of DAGs are representative of the behavior of the time-dependent nonlinear multivariate interactions. Our results show the emergence of patterns of persistent behavior (Figure 3.5). Both during a clear sky day and during the action of a solar eclipse, the dynamics maintain a certain order on the causal structure that exhibits particular attributes based on the cluster, whose effects persist for a period to create patterns. We observe that there is a certain stability in the causal structure of interactions that allows for reorganization as the atmospheric conditions change, even during the effect of a strong intervention through the eclipse. This lead us to postulate that while each variable can tell us about its own temporal dynamics (Figures 3.2c and d, respectively), the high frequency observations also carry information about the influence of other variables on their dynamics, which together, sustain the system in dynamics. The slight reduction in the connectivity of the multivariate system where the wind dynamics tend to decouple from the thermodynamics, and from CO_2 and H_2O dynamics (Figure 3.7, cluster #2), have a strong pattern that lasts longer in comparison to when the system experience a larger interdependency (Figure 3.7, cluster #3, in yellow, and cluster #1, in blue, respectively). Our results suggest that at turbulence scales the system tends to alternate states of increased and decreased connectivity but bounded within a range of general behavior.

3.4.2 A novel approach for exploring the structural behavior

Our approach for exploring the evolution of the causal structure of a complex system can be inquired with respect to the law of motion. This discussion section is inspired by Marletto [132]. A dynamical law is a description of where a system goes given that its motions start at a certain point in space and time. In our case, in the absence of a causal model able to predict the behavior of the dynamical system, we estimate the evolution of the structure causal model using a sequence of DAGs of the near-instantaneous dynamics of turbulence. Each of those DAGs represents the dynamic state of the connectivity of the multivariate system at consecutive times on the overall causal trajectory of the system. Such behavior emerges from the nonlinearities within the system, where the interdependencies among components of the system change over time and are not imposed. Therefore, we do not have yet a dynamical causal law that could provide the recipe to replicate the behavior of turbulence dynamics in terms of its connectivity with information flow. The

laws of motion of a dynamical system are able to describe the trajectory of a system with the property of reversibility, meaning that after the sequence is described, it is possible to go back without any uncertainty because there is only one predecessor. In our case, a causal dynamical law might encounter bifurcations occurring along the trajectory, then it wont be possible to go back and reconstruct the sequence given the initial conditions of the system. Therefore, in the absence of a causal law, we proposed here a different approach to explore the evolution of the structure of the dynamics of the system, using a structure causal model to determine the patterns of similar connectivity dynamics of a complex multivariate system.

In addition, the study of the evolution of the structure causal model may raise questions regarding how the connectivity of the system may "look like" typically and under certain interventions. Thus, we have observed that the variables in turbulence coexist and inform each other in a persistent re-organizational state around a range of dynamical behavior instead of being in a fixed configuration. The connectivity of the complex system fluctuates within a range of variations, allowing us to define patterns of evolution of its dynamics. Under the effects of a rapid intervention such as a solar eclipse, which is able to perturb the dynamics of the whole system, the connectivity defined by the number of links that allow information flow within the system decreases momentarily and eventually reaches a new dynamical state where the connectivity is restored. While the connectivity of some variables remain the same across DAGs, others are temporarily diminished or even suppressed as the system responds to the intervention. Here we define this as the "system resilience to information flow", that reflects the ability of a system to continue functioning when there has been a disruption or intervention in the typical behavior of the system to ensure the continuity of information flow.

3.5 Conclusions

We use the time series of observations at 10Hz from a 25m tall eddy covariance flux tower to explore the evolutionary dynamics of the system at turbulence scales. Non-linear interactions among components in a multivariate complex system determine its causal structure, which evolves due to changes in information flow. The causal structure refers to a network representation that shows the connectivity of the information transfer among components of the system. In this study, we explore the evolution of the causal structure of turbulence during a clear sky day and under the effects of a solar eclipse, a unique natural intervention that is able to intervene the dynamics of the whole complex system.

Our novel approach to explore the evolution of the causal structure of a complex system using DAGs allows us to investigate the rapid interdependencies of turbulence at the biosphere-atmosphere interface. By using 10Hz data, we have moved beyond the detection of contemporaneous dependencies observed in time-averaged datasets [110], to explore the lagged influences resulting from self-feedback and cross-dependencies that shape the causal structure of turbulence. Our approach is based on information-theoretic metrics and a k-means type of clustering approach to

classify the DAGs in groups of similar and unique causal structures. When we estimate the evolution of the causal structure of the turbulence system under the influence of the rapid but progressive solar radiation limitation and recovery experienced during a solar eclipse, our results show the formation of behavioral patterns of similar causal structure. Our findings have broad implications in the understanding of turbulence systems. We find three well-defined clusters of structural connectivity in the system.

Our results show that the system is likely to experience a temporary and persistent reduction in its connectivity, which is eventually recovered to a higher state of connectivity. We also find the peculiar cases of the decrease of self-feedback of its dynamics, such as for T and CO_2, which are addressed from methodological and phenomenological perspectives. These analyzes lead us to postulate that turbulence at the biosphere-atmosphere interface coexists in a persistent reorganizational state in terms of its connectivity, which reaches a new behavioral state after the effect of a rapid intervention that perturbs the dynamics of the whole system passes. Such an ability of turbulence to remain in a range of the organizational state of its causal structure that sustains the flow of information is defined in this study as "system resilience to information flow".

Inferring causality to draw conclusions from observations requires some assumptions [10], [57], such as the Markov assumption, faithfulness of the DAGs to represent the dynamics of the system, causal sufficiency, causal stationarity inside each 3-minute window, and no contemporaneous causal effects. We anticipate that the study of the causal structure of the dynamics of the connectivity of complex systems at high frequency will contribute to the exploration of the dynamics of the information flow for more processes at the turbulence scale and over longer time periods to inform Earth system modelling efforts and prediction.

Chapter 4

Causal interaction in high frequency turbulence at the biosphere-atmosphere interface: Structure - function coupling

4.1 Introduction

[1]

"Does structure determine function, or does function determine structure, or does each determine the other, and if the latter, how can it be?" [133]

High frequency data at the biosphere-atmosphere interface encode the dynamics of a complex and open dissipative system where processes of turbulence occur. In this complex system, multiple variables such as horizontal and vertical wind speed (WS, U_z), air temperature (T), water vapor (H_2O), and carbon dioxide (CO_2), interact through nonlinear dynamics. In 3 Hernandez Rodriguez and Kumar, we use high frequency data to reveal the causal structure of dynamic interdependencies among these components in turbulence. Such dependencies arise as fluctuations in one variable at any time, instigating variations in other linked variables at a future time [49], creating a causal structure of interdependencies represented by directed acyclic graphs (DAGs) [10]. In a DAG, the edges linking lagged states of variables, denoted as nodes are used to create paths for information flow in a forward direction in time, which can be seen as dynamic channels for communication between a set of lagged sources of information, named "parents", influencing the current state of a target variable [9], [50]. We combined information-theoretic measures to estimate causal influences using a Granger approach [117]. We explored the evolution of the causal structure of turbulence for a range of conditions, during a clear

[1]This chapter is in review for Journal of Geophysical Research [Hernandez Rodriguez and Kumar, 2022]

sky day and during a natural intervention on the whole system dynamics resulting from a solar eclipse. We used high frequency data from an eddy covariance flux tower, located in an intensively managed agricultural landscape in the Midwestern US. We estimated a series of consecutive DAGs over non-overlapping windows using Tigramite algorithm [43], [51]–[53], which are based on information theory metrics [109]. We then used a distance-based classification in conjunction with a k-means clustering approach to cluster the DAGs with similar structure reflecting causal dependencies. Our results show that the system is likely to experience a temporary and persistent reduction in its connectivity, which is eventually recovered. A major finding is the temporal "decoupling" of variables, meaning a significant reduction of self-dependence and cross-connectivity, such as for T and CO_2 [134]. These led us to postulate that turbulence at the biosphere-atmosphere interface exists in a persistent reorganizational state, which reaches a new behavioral state after the effect of a rapid intervention that perturbs the dynamics of the whole system passes [134]. Therefore, we proposed the concept of "system resilience to information flow" Hernandez Rodriguez and Kumar [134], which refers to the ability of a system to remain in a stable range of the organizational state of its causal structure that sustains the flow of information, even under the effect of an intervention.

From the study of the evolution of the causal structure, we found how information is encoded in the dynamics that arise from interdependencies among components of the system. Does the evolution of the causal structure reflect different underlying dynamics? To answer this question we analyze the representative dynamics of the clusters using the DAG nearest to the cluster centroid. Given that each of those representative DAGs is formed by edges linking the lagged variables through which information flows, we ask, what do changes in information flow over time tell us? Here we aim to explore the relationship between the causal structure and the associated function of the system. We argue that the function of the system is the ability to sustain its dynamics through information flow among its multiple components. We aim to explore if the information flow is changing in time, is it because (1) functional differences are a reflection of structural differences, or (2) although there are structural differences, there are no functional differences. Here we hypothesize that functional outcomes arriving from interactions among component variables are a result of information flow and the associated structure of connectivity between them, which places constraints on and supports the processes [49].

To test our hypothesis, we build upon our causal structure analysis [134] to estimate the information flow in each DAG cluster using a multivariate causal history approach [9], [54]. The information transfer from the causal history of a multivariate system, developed by Jiang and Kumar [50], accounts for all interacting variables through the entire time history of the system when influencing the outcome of any variable at the current time, t. Using a causal history approach, we characterize the joint influence of lagged self- and cross-dependencies in determining the current state of each variable [9], [50]. While self-dependency refers to how a variable's own history influences its present state, cross-dependency refers to the influence of all other variables arising through interactions as represented in a DAG [9]. Using a varying partitioning time lag τ_C, the causal history can be partitioned into immediate and distant causal

histories, which can be further partitioned into self- and cross-dependencies [54]. Further more, these interactions can be partitioned and classified as unique, synergistic, and redundant (SUR) information flow, through the partial information decomposition (PID) approach [47], [48] using momentary partial information decomposition (MPID) [54]. Using this framework, we estimate the contribution of each information type, i.e., synergistic, unique, and redundant for each individual target variable in the system. We also determine whether the dynamics of turbulence exhibit long- or short-term memory, and how the influence of the self- or cross-feedback interactions from the immediate and distant causal history sustain the short- and long-term dynamics of the system.

While a few previous studies of causality in turbulence have mostly focused on transfer entropy [65]–[68], we propose that the use of a SUR framework that uses MPID provides a more informative perspective of the underlying dynamics of the system, in terms of the type and amount of information flowing among the system components over time. To answer our research question, we further compare our results between the different clusters of DAGs, to determine if the different causal structures as captured by the representative DAGs in each cluster reflect differences in their self-organized dependencies and associated information flow. Based on our results, we further discuss the relationship between interactions in the causal history of a variable and the observed long-range dependency as exhibited through spectral analysis.

The chapter is organized as follows: Following this introduction, in Section 4.2, we summarize the concepts and metrics to estimate the information flow in the SUR framework. In Section 4.3, we present the results of the causal history analysis for each of the clusters found in the exploration of the evolution of the causal structure of turbulence [134]. Section 4.4 discusses the relationship between the evolution of the causal structure and the corresponding functionality of the system. Section 4.5 concludes the chapter while mentioning the potential implementation of our findings and future work.

4.2 Characterizing causal dependency

4.2.1 Quantifying information

Here we review the basic information-theoretic measures based on Shannon's entropy [109] used to quantify the information flow from nonlinear dependencies among multiple variables. The uncertainty of a dynamic variable X_t is quantified by *Shannon's entropy*:

$$H\left(X_t\right) = -\sum_{x_t \in X_t} p\left(x_t\right) \log p\left(x_t\right), \tag{4.1}$$

where $p\left(x_t\right)$ is the probability of X_t. Now, if we consider a variable Y_t, the *conditional entropy* accounts for the

uncertainty of X_t that remains given the knowledge of another variable Y_t:

$$H\left(X_t \mid Y_t\right) = -\sum_{x_t \in X_t, y_t \in Y_t} p\left(x_t, y_t\right) \log \frac{p\left(x_t, y_t\right)}{p\left(x_t\right)}, \tag{4.2}$$

where $p(x_t, y_t)$ corresponds to the joint probability of X_t and Y_t. The *mutual information* accounts for the dependency between X_t and Y_t, symmetrically measuring the reduction of uncertainty of one variable given the knowledge of the other:

$$\begin{aligned} I\left(X_t; Y_t\right) &= \sum_{x_t \in X_t, y_t \in Y_t} p\left(x_t, y_t\right) \log \frac{p\left(x_t, y_t\right)}{p\left(x_t\right) p\left(y_t\right)} \\ &= H\left(X_t\right) - H\left(X_t \mid Y_t\right) = H\left(Y_t\right) - H\left(Y_t \mid X_t\right). \end{aligned} \tag{4.3}$$

Given an additional variable Z_t, which is influenced by the union of X_t and Y_t, the mutual information takes the form of $I(Z_t; X_t, Y_t)$. Further, the *partial information decomposition* (PID) [48] decomposes this total information shared with Z_t from X_t and Y_t into three components. The components of PID are: *Synergistic information*, S, which is the information jointly provided by X_t and Y_t; *Redundant information*, R, which is the overlapping information between X_t and Y_t; and, *unique information*, U_X and U_Y, which is the information provided by X_t and Y_t individually. They satisfy:

$$I\left(Z_t; X_t, Y_t\right) = S + R + U_X + U_Y \tag{4.4}$$

In other words, the SUR framework decomposes the mutual information between the three variables. To estimate PID for time series data, we use the approach developed by Goodwell and Kumar[47].

4.2.2 Estimating information flow using DAGs

The causal structure of a dynamical system is represented as a DAG [10], a structural causal model of the dynamics of a multivariate complex system. We consider that the main function of the causal structure is a cause and a consequence of information flow among interacting variables. Here, we describe the information-theoretic measures used to estimate information flow using DAGs for multivariate time series [9], that represent the temporal dependencies of the system. In a multivariate complex system with N variables, $\vec{X}_t = \{X_t, Y_t, Z_t, \ldots\}_N$, varying in time t, the current state of any variable $Z_t \in \vec{X}_t$ is the result of all earlier interactions of its dynamics [9], [50]. Such prior dynamics are called causal history, denoted as $\vec{X}_t^- = \left\{\vec{X}_{t-1}, \vec{X}_{t-2}, \vec{X}_{t-3}, \ldots\right\}$ [9], [50]. We estimate information flow using a DAG representation for time series, $G = \left(\vec{X}_{t+1}^-, E\right)$, where each node refers to the state at time t of a variable in $\vec{X}_t$ and E is the collection of edges. If a directed edge E links two nodes $Y_{t-\tau}$ and Z_t, its is denoted by $Y_{t-\tau} \rightarrow Z_t$, where τ is a positive time lag. The parent set of a target variable at time t, Z_t, is denoted as $P_{Z_t} \equiv \left\{Y_{t-\tau} : Y_t \in \vec{X}_t, \tau > 0, Y_{t-\tau} \rightarrow Z_t\right\}$ [9], [50]. A node $Y_{t-\tau}$ can also

be linked to a target indirectly through a causal path $C_{Y_{t-\tau}\to Z_t}$ which is a set of nodes connected by a sequence of edges linking from $Y_{t-\tau}$ to Z_t, that is, $C_{Y_{t-\tau}\to Z_t} = \left\{V_{t-\tau_V} : V_t \in \vec{X}_t, \tau_V > 0, Y_{t-\tau} \to \ldots \to V_{t-\tau_V} \to \ldots \to Z_t\right\} \cap \{Y_{t-\tau}\}$. To estimate the information flow from one or multiple sources on the current state of a variable, we use one of the following possibilities: through a direct edge (Figure 4.1a) [43], by one (Figure 4.1b) [52] or two causal paths (Figure 4.1c) [54], or by the whole causal history (Figure 4.1d) [9], [50].

Information flow through a directed edge between two nodes, $Y_{t-\tau}$ and Z_t, denoted by $Y_{t-\tau} \to Z_t$ (Figure 4.1a), exist only if a perturbation in $Y_{t-\tau}$ propagates to Z_t, and it is measured as the *conditional mutual information* (CMI) [43],

$$I\left(Z_t; Y_{t-\tau} \mid \vec{X}_t^- \backslash Y_{t-\tau}\right) \geq 0 \tag{4.5}$$

where $\backslash$ is the exclusion symbol, such as $\vec{X}_t^- \backslash Y_{t-\tau}$ represents all the nodes in the causal history excluding $Y_{t-\tau}$. CMI estimates the information flow from a lagged variable $Y_{t-\tau}$ to the current state of a target variable Z_t, conditioned on the knowledge of the rest of the dynamics of all interacting variables in the causal history $\vec{X}_t^-$ excluding $Y_{t-\tau}$ [9]. However, the computation of CMI is infeasible due to the potential infinite number of nodes arising from the causal history (i.e. curse of dimensionality). To make the computation possible, the *momentary information transfer* (MIT) was developed [43] using the Markov property for DAG [123], which states that a node Z_t is statistically independent of the rest of the history if its parents P_{Z_t} are given, where $P_{Z_t} = \left\{X_{t-\tau} : X_{t-\tau} \in \vec{X}_{t-\tau}, \tau > 0, X_{t-\tau} \to Z_t\right\}$ [9], [50]. MIT quantifies the direct interaction between two nodes, $Y_{t-\tau}$ and Z_t, acting as source and target, respectively, excludes any information from other nodes that may be flowing through the source or directly to the target [9]. MIT is given by:

$$I_{Y\to Z}^{MIT}(\tau) = I\left(Z_t; Y_{t-\tau} \mid P_{Z_t}, P_{Y_{t-\tau}} \backslash P_{Z_t}\right) \tag{4.6}$$

In addition to a direct influence through an edge, *information flow through a causal path* (Figure 4.1b) occurs when a lagged source node $Y_{t-\tau}$ indirectly affect a target node Z_t. The causal path is denoted as $C_{Y_{t-\tau}\to Z_i}$. The implementation of the Markov property for information flow through a causal path is called *momentary information transfer along causal path* (MITP) [51], and is given by:

$$I_{Y\to Z}^{MITP}(\tau) = I\left(Z_t; Y_{t-\tau} \mid P_{Z_t}, P_{C_{Y_{t-\tau}\to Z_t}} \backslash P_{Z_t}\right) \tag{4.7}$$

where $P_{C_{Y_{t-\tau}\to z_t}}$ is the parent set of the causal path $C_{Y_{t-t}\to Z_t}$.

As an extension of the MITP, *information flow through two causal paths* (Figure 4.1c) estimates the influence of two lagged sources $X_{t-\tau_X}$ and $Y_{t-\tau_Y}$ on a target Z_t, through the corresponding causal paths $C_{X_{t-\tau_X}\to Z_t}$ and $C_{Y_{t-\tau_Y}\to Z_t}$,

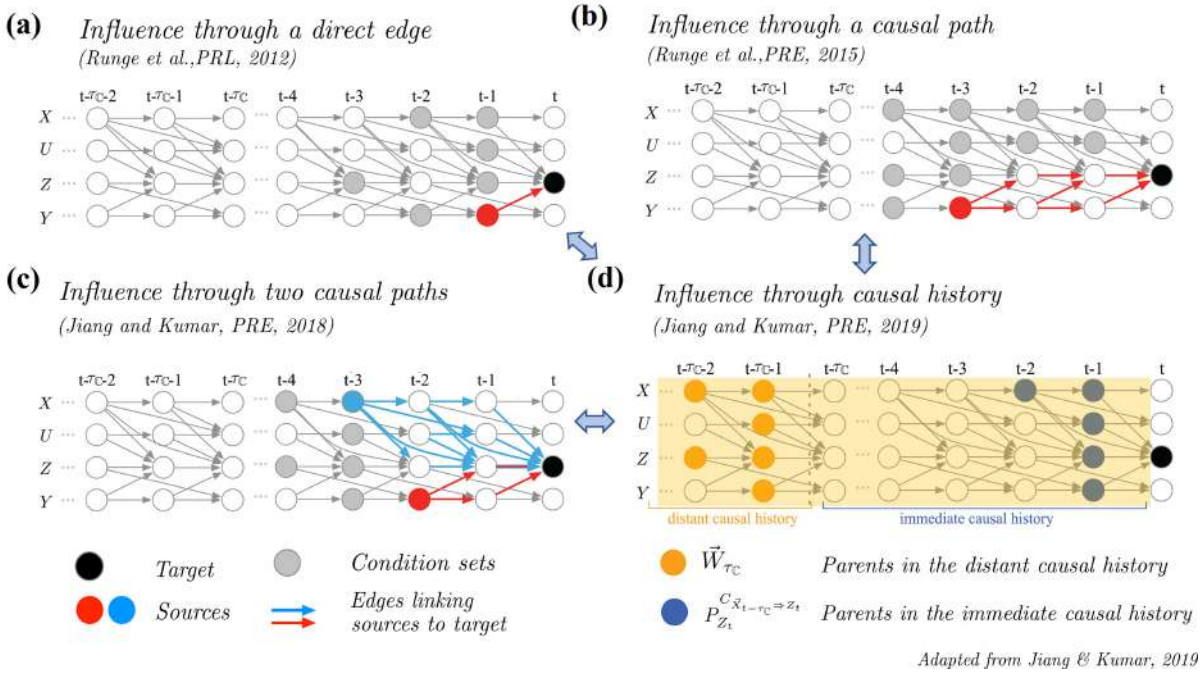

Figure 4.1: Illustration adapted from Jiang and Kumar [9] that shows the information flow to a target node Z_t in a quadvariate complex system from (**a**) Y_{t-1} through a directed edge; (**b**) Y_{t-3} through the corresponding causal path $C_{Y_{t-3}} \rightarrow Z_t$; (**c**) X_{t-4} and Y_{t-2} through the union of the corresponding two causal paths $C_{X_{t-4};Y_{t-2}} \rightarrow Z_t$; and (**d**) the entire causal history, which can be partitioned into immediate and distant causal histories based on a partition time lag $\tau_{\mathbb{C}}$.

respectively [54], as the *conditional mutual information*:

$$I\left(Z_t; X_{t-\tau_X}, Y_{t-\tau_Y} \mid P_{Z_t}, P_{C_{\{X_{t-\tau_X}, Y_{t-\tau_Y}\}\Rightarrow Z_t}} \backslash P_{Z_t}\right) \tag{4.8}$$

where $C_{\{X_{t-t_X}, Y_{t-\tau_Y}\}\Rightarrow Z_t} = C_{X_{t-\tau_X}\to Z_t} \cup C_{Y_{t-\tau_y}\to Z_t}$ is a subgraph consisting of the union of the causal paths and $P_{C_{\{X_{t-\tau_X}, Y_{t-\tau_Y}\}\Rightarrow Z_t}}$ refers to the parent set [9]. When the influence through two causal paths affect a target, the interaction between these two causal paths can be characterized through PID in what is called *momentary partial information decomposition* (MPID) [54]:

$$I\left(Z_t; X_{t-\tau_X}, Y_{t-\tau_Y} \mid P_{Z_t}, P_{C_{\{X_{t-\tau_X}, Y_{t-\tau_Y}\}\Rightarrow Z_t}} \backslash P_{Z_t}\right) = S_c + R_c + U_{X,c} + U_{Y,c} \tag{4.9}$$

where the subscript c in the right hand side represents that the PID is associated with causal paths [50]. MPID accounts for the information only going through the pathways linking the sources and the target, with the influence from earlier dynamics excluded through conditioning [50].

The state of a target variable Z_t is the result of the prior states of all interdependent variables in the system, i.e. *causal history* $\vec{X}_t^-$, with information flowing through a multitude of different pathways in the DAG (Figure 4.1d) [50]. The *total information flow* ($\mathscr{T}$) from the causal history can be partitioned into two different complementary components. The immediate causal history arising from all the previous states from time step $t-1$ up to the time step $t-\tau_{\mathbb{C}}$, called *immediate causal history* ($\mathscr{J}$), and the remaining earlier dynamics, called *distant causal history* ($\mathscr{D}$), respectively (Figure 4.1d) [50], can be estimated as follows:

$$\mathscr{T} = I\left(Z_t; \vec{X}_t^-\right) = \mathscr{J}\left(\tau_{\mathbb{C}}\right) + \mathscr{D}\left(\tau_{\mathbb{C}}\right) \tag{4.10}$$

with

$$\mathscr{J}\left(\tau_{\mathbb{C}}\right) = I\left(Z_t; C_{\vec{X}_{t-\tau_{\mathbb{C}}}\Rightarrow Z_t} \mid \vec{X}_t^- \backslash C_{\vec{X}_{t-\tau}\Rightarrow Z_t}\right) \tag{4.11a}$$

$$\mathscr{D}\left(\tau_{\mathbb{C}}\right) = I\left(Z_t; \vec{X}_t^- \backslash C_{\vec{X}_{t-\tau_{\mathbb{C}}}\Rightarrow Z_t}\right) \tag{4.11b}$$

where, immediate ($\mathscr{J}$) and distant ($\mathscr{D}$) causal histories are the partitioning of the total information $\mathscr{T}$ at time lag $\tau_{\mathbb{C}}$. Further, the Markov property of the DAG provides the feasibility of computing equations (4.11) by simplifying $\mathscr{J}(\tau_{\mathbb{C}})$ and $\mathscr{D}(\tau_{\mathbb{C}})$ as [50]:

$$\mathscr{J}\left(\tau_{\mathbb{C}}\right) = I\left(Z_t; P^{C}_{Z_t}\vec{X}_{t-\tau_{\mathbb{C}}\Rightarrow z_t} \mid \vec{W}_{\tau_{\mathbb{C}}}\right) \tag{4.12a}$$

$$\mathscr{D}\left(\tau_{\mathbb{C}}\right) = I\left(Z_t; \vec{W}_{\tau_{\mathbb{C}}}\right) \tag{4.12b}$$

where $\vec{W}_{\tau_{\mathrm{C}}}$ refers to the total information from the distant causal history, quantified as $\mathscr{D}$ in equation (4.12b).

Furthermore, the analysis of multivariate time series using DAG representation allows for the partitioning of the immediate and distant causal histories of any target node $Z_t \in \vec{X}_t$ of a system into self- and cross-dependencies [9]. Self-dependency refers to the variable's own history influencing its present state, denoted by $\vec{Z}_{\mathscr{D}} \equiv \left\{Z_{t-\tau} : Z_{t-\tau} \in \vec{W}_{\tau_C}\right\}$. Cross-dependency refers to the influence of all other variables arising through interactions in a DAG, denoted by $\vec{Z}'_{\mathscr{D}} \equiv \vec{W}_{\tau_{\mathrm{C}}} \backslash \vec{Z}_{\mathscr{D}}$. Using the PID framework, the interactions in the distant causal history $\mathscr{D}$ between a target and other variables can be quantified as a function of the partitioning time lag $\tau_{\mathbb{C}}$ as follows:

$$\mathscr{D}\left(\tau_{\mathbb{C}}\right) = I\left(Z_t; \vec{W}_{\tau_{\mathrm{C}}}\right) = I\left(Z_t; \vec{Z}_{\mathscr{D}}, \vec{Z}'_{\mathscr{D}}\right) \tag{4.13a}$$

$$= S_{\mathscr{D}}\left(\tau_{\mathbb{C}}\right) + R_{\mathscr{D}}\left(\tau_{\mathbb{C}}\right) + U_{\text{self},\mathscr{D}}\left(\tau_{\mathbb{C}}\right) + U_{\text{cross},\mathscr{D}}\left(\tau_{\mathbb{C}}\right) \tag{4.13b}$$

where $S_{\mathscr{D}}$ and $R_{\mathscr{D}}$ are the synergistic and redundant information from distant causal history, respectively, and $U_{\text{self},\mathscr{D}}$ and $U_{\text{cross},\mathscr{D}}$ are the unique information from the self- and cross-dependencies, respectively. Likewise, for the immediate history, $\mathscr{J}$, the partitioning is given by

$$\mathscr{J}\left(\tau_{\mathbb{C}}\right) = I\left(Z_t; P^{C}_{Z_t}\vec{X}_{t-\tau_C} \Rightarrow Z_t \mid \vec{W}_{\tau_C}\right) = I\left(Z_t; \vec{Z}_{\mathscr{J}}, \vec{Z}'_{\mathscr{J}} \mid \vec{W}_{\tau_{\mathrm{C}}}\right) \tag{4.14a}$$

$$= S_{\mathscr{J}}\left(\tau_{\mathbb{C}}\right) + R_{\mathscr{J}}\left(\tau_{\mathbb{C}}\right) + U_{\text{self},\mathscr{J}}\left(\tau_{\mathbb{C}}\right) + U_{\text{cross},\mathscr{J}}\left(\tau_{\mathbb{C}}\right) \tag{4.14b}$$

Therefore, from equations (4.12) to (4.14), we can estimate the PID of the information from the entire causal history, $\mathscr{T}$, as follows:

$$\mathscr{T} = I\left(Z_t; P_{Z_t}\right) = \mathscr{J} + \mathscr{D} \tag{4.15a}$$

$$= S_{\mathscr{J}} + R_{\mathscr{J}} + U_{\text{self},\mathscr{J}} + U_{\text{cross},\mathscr{J}} + S_{\mathscr{D}} + R_{\mathscr{D}} + U_{\text{self},\mathscr{D}} + U_{\text{cross},\mathscr{D}} \tag{4.15b}$$

$$= S_{\mathscr{T}} + R_{\mathscr{T}} + U_{\text{self},\mathscr{T}} + U_{\text{cross},\mathscr{T}} \tag{4.15c}$$

We use the rescaled redundancy approach of Goodwell and Kumar [47] incorporated by Jiang and Kumar [9], [50] for computing the PID for $\mathscr{D}$, $\mathscr{J}$, and $\mathscr{T}$, where the redundant information is estimated by considering the mutual dependency between two sources and ensures a non-negative information partitioning [9]. We compute $\mathscr{D}$ and $\mathscr{J}$, along with their PIDs, using equations (4.14) and (4.13), respectively. Then, the PID of $\mathscr{T}$ is obtained based on the sum of $\mathscr{D}$ and $\mathscr{J}$ according to equation (4.10).

This review of information flow metrics has summarized how to quantify the dynamics that sustain the whole multivariate system [9]. Although, two important aspects should be considered when implementing the previous information flow metrics. First, the dimensionality can be high for reliable estimations when computing $\mathscr{J}$, $\mathscr{D}$, and PID, even after using the Markov property for DAG [43]. The dimensionality grows as the number of variables increases and/or as the number of lags that influence a target increases [9]. Therefore, we use the *momentary information weighted transitive reduction* (MIWTR) proposed by Jiang and Kumar [9], which reduces the dimensionality of the DAG under analysis. Briefly, MIWTR removes edges linking $\vec{W}_{\tau_C}$ with the immediate causal history, $C_{\vec{X}_{t-\tau_c} \Rightarrow Z_t}$ and then exclude the nodes $\vec{W}_{\tau_C}$ not directly linked to $C_{\vec{X}_{t-\tau_C} \Rightarrow Z_t}$, to obtain a reduced $\vec{W}_{\tau_C}$. The details of the MIWTR technique are described in Jiang and Kumar [9]. Second, the adequate estimation of information metrics requires a considerable data length. Jiang and Kumar [9] found that the cardinality reduction based on MIWTR does not affect the estimation of information-theoretic measures significantly when the time series data length is sufficient. Data length greater than 1000 is considered as the minimum data length sufficient to achieve reliable estimations [9].

4.3 Results

We explore the relationship between causal structure and associated function, using the representative DAGs in each cluster determined in Hernandez Rodriguez and Kumar [134] (Figure 4.2). We define functionality as the nature of interactions as discerned through the SUR components. Our aim is to discern the structure-function duality in the causal interactions that shape the observed dynamics in the causal history. If the different causal structures give rise to different behavior in their information flow, then we may infer that the functional differences are a reflection of structural

differences. Otherwise, we may infer that different structural behaviors may give rise to similar functional behavior despite the structural differences, indicating that there are no differences in terms of the ability of the system to sustain its dynamics by transferring information among its components.

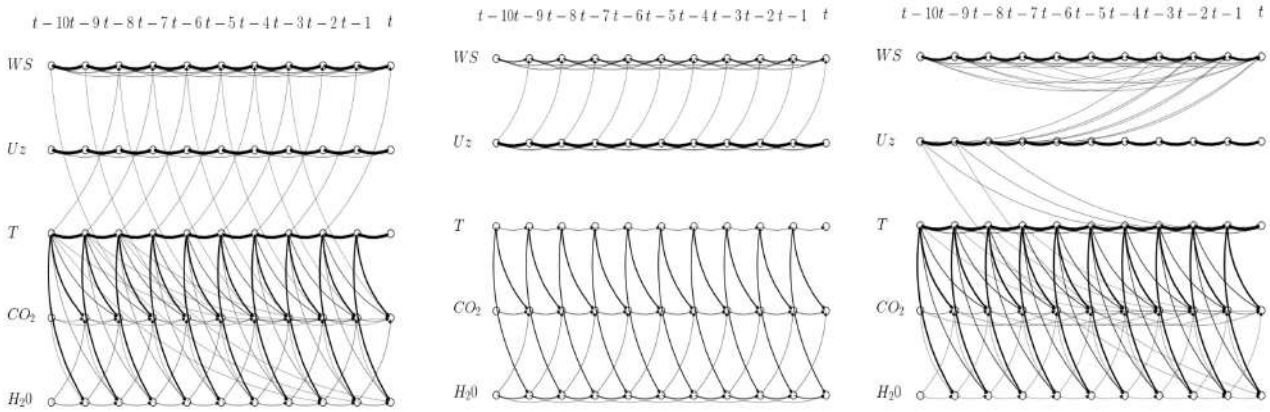

Figure 4.2: DAGs from the observed time series that are closest to each cluster's centroids show the average of the dynamics of (a) cluster #3, (b) cluster #2, and cluster #1. The DAGs for time series are estimated using Tigramite algorithm [43], [51]–[53].

4.3.1 Information flow to a target across DAG clusters

To illustrate the approach described above, we consider CO_2 and H_2O as examples. To estimate their current state (Figure 4.3 left and right column, respectively), we use the causal history approach [50], [54], using the momentary information weighted transitive reduction (MIWTR) [9]. We select a varying partitioning time-lag τ_C to partition the influence of the historical dynamics of all interacting components into immediate and distant history, respectively. By changing τ_C we can then gauge how the influence of immediate history declines or continues to persist through the time history. Furthermore, we explore the differences in self- and cross-dependencies influencing the current state of a target variable in its immediate and distant history. As stated earlier, self-dependence refers to the variable's own history influencing its present state, cross-dependence refers to the influence of all other variables arising through interactions as represented in the DAG (see equations 4.15).

In Figure 4.3, we compare the influences across clusters using the DAGs for when $\tau_C = 10$ on the current state of CO_2 and H_2O. That is, the immediate history accounts for the influence of the parents up to one second in the history of the variable. These results show that the distant history of the vertical wind speed, U_z, influence the current state of both CO_2 and H_2O, and it is specially noticeable for clusters #3 and #1. In general, we observe that both self- and cross-dependencies are lower in cluster #2 in comparison to the other two clusters. We also observe that the horizontal wind speed, WS, does not influence the current state of CO_2 and H_2O for clusters #1 and #2. These results suggest that

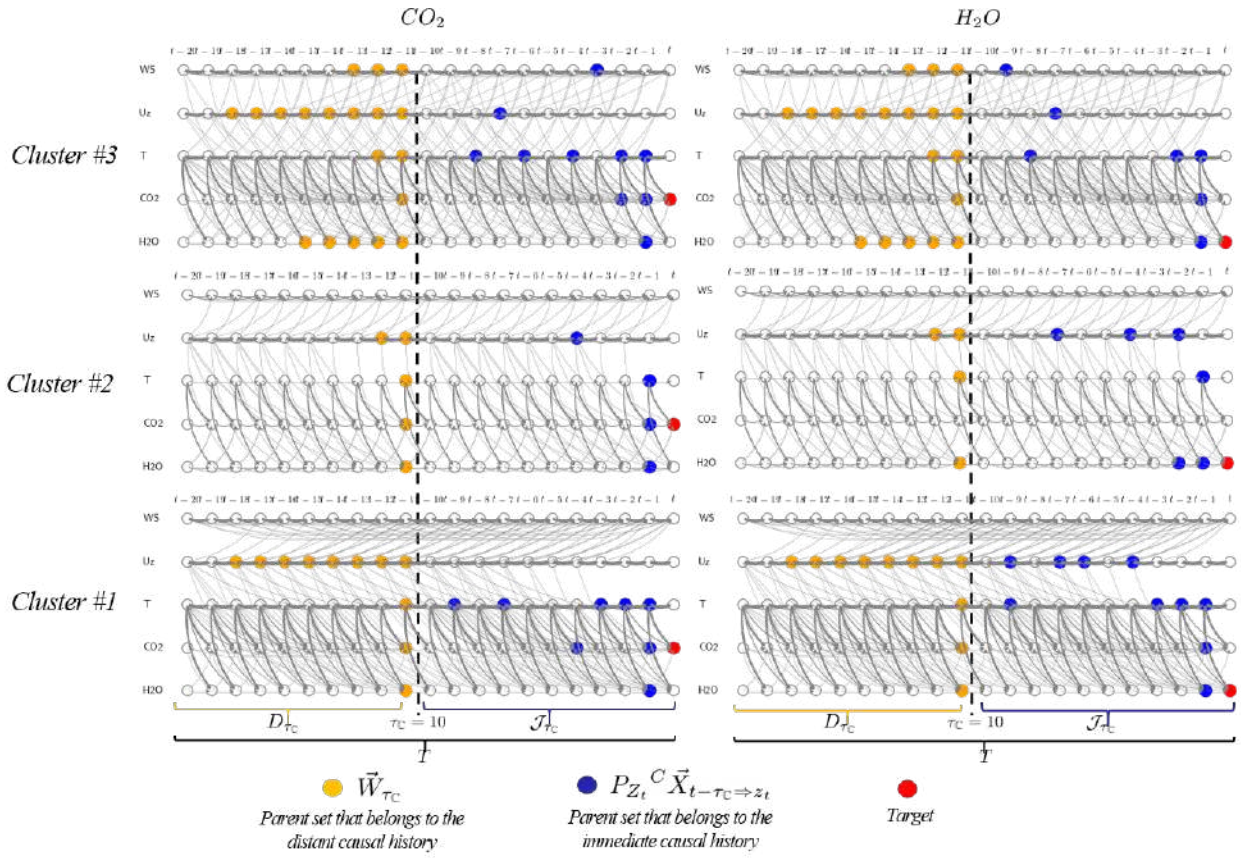

Figure 4.3: Partitioned causal history for directed acyclic graphs (DAG) using the causal history approach [50], [54] with MIWTR [9] for the present state of CO_2 (left column) and H_2O (right column). The DAGs correspond to the representative time series that are closest to the centroid of each cluster found in the causal structure analysis [134]. The nodes refer to the state of the variables at times $t - \tau$ (as marked on top), where τ is a positive time-lag, here equal to 0.1s (i.e since data is at 10Hz). We set the partitioning time $\tau_C = 10$ in this illustration to divide the influence of the causal history based on its immediate and distant history on the current state of the target variable. The nodes influencing the target (red color solid nodes) can come from the parent set in the immediate (blue color solid nodes) or distant (yellow color solid nodes) history.

the stagnant environment characterized by low horizontal wind speed results in turbulence decoupling of the scalar transport from the dynamics [135], and it is reflected as the absence of information flow from WS on the dynamics of other variables (Figure 4.2, clusters #1 and #2). These results are consistent with observations noted during total solar eclipses regarding the slackening of wind speed due to the boundary layer stabilization [135], [136]. A time-shift between the time of totality and the lagged response of ecohydrological variables is also a recurrent observation [137]. It is observed that the stagnant environment with minimal wind speeds shows turbulence decoupling the canopy and atmosphere, which produced uncertainty when estimating transpiration rates [135]. While the effect of the absence of incoming solar radiation on the longwave upward radiation is lagged only for few minutes, the fluxes that depend on surface temperature or photosynthesis, such as latent and sensible heat, and CO_2 flux, responded to the change of the net radiation by reducing their magnitudes, and then were restored to their typical values after the increase of turbulence [135], [137]. For a clear sky day behavior or under the effect of an intervention on the whole system dynamics, such as a solar eclipse, the analysis of changes in the structure of information flow across multiple variables helps us to characterize how the functionality of the ecohydrological system at a turbulence scale evolves. To determine the relationship between the causal structure and function of the system, we determine attributes of the interactions among components of the system.

4.3.2 Characteristics of information flow across clusters

After finding the lagged sources of information to the current state of a target variable and for a singular partitioning time lag τ_C, we now estimate the amount of information flowing from the causal history of each variable for multiple τ_C. We estimate the maximum information from the entire causal history, $\mathscr{T}$, which is invariant with respect to the partitioning time lag τ_C and the percentage of information from the distant history as a function of the partitioning time lag τ_C (Figure 4.4). Typically, for small τ_C a large fraction of causal information comes from the distant history, $\mathscr{D}$, and as τ_C increases, the information from $\mathscr{D}$ decreases while for information from the immediate history $\mathscr{I}$ increases. Furthermore, after the initially high $\mathscr{D}$, the percentage of information from distant history with respect to the total information, $\mathscr{D}/\mathscr{T}$, typically exhibits a decrease, excepting for the vertical component of the wind velocity, U_z (Figure 4.4). For U_z as τ_C increases, the signal quickly dies out, which suggests that the percentage of information coming from its distant history is small. Therefore, we can say that the vertical wind velocity, U_z, highly relies on its short immediate history, which is characteristic of a short-term memory process.

Overall, as τ_C increases, $\mathscr{D}$ accounts for less than 50% of $\mathscr{T}$, suggesting that processes in turbulence are mostly sustained by short-term causal interactions (Figure 4.4). Across clusters, we observe that the magnitude of the ratio between $\mathscr{D}$ and $\mathscr{T}$ changes depending on the variable. This suggest that the causal structure of the clusters contains information about the short- and long-term interdependencies among variables, which imply that the differences in

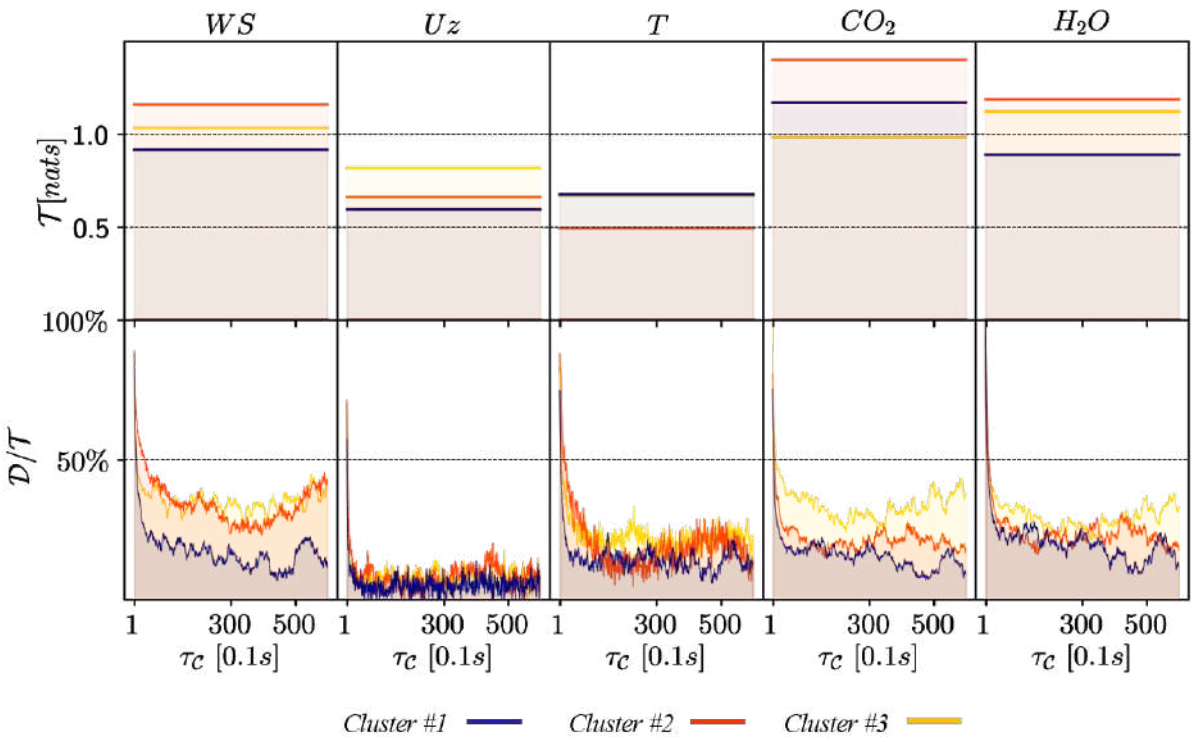

Figure 4.4: Information flow from the total causal history, $\mathscr{T}$ (top), and the ratio between the distant, $\mathscr{D}$, and the total causal histories (bottom), over the partitioning time lag τ_C for the cluster #1 (blue line), cluster #2 (red line), and cluster #3 (yellow line) based on the directed acyclic graphs for time series in Figure 4.2. Notice that τ_C takes values from 1 to 600, accounting for influences coming from up to one-minute.

information flow among variables reflects how the process may interact in the causal structure.

4.3.3 A SUR framework to infer characteristics of multivariate turbulence

Do the self-or cross-dependencies dominate the immediate and distant dynamics of each process? Knowing how the multivariate system informs the short-term dominated processes at a turbulence scale, it would provide insights into the oscillatory behavior observed for temperature T in cluster #2 (Figure 4.4 red line in T panel at the bottom), which might be related to a periodic component as a function of τ_C. To further investigate the self-and cross-dependencies for each variable and across clusters, here we use a partial information decomposition (PID) approach [47], [48] implemented in the causal history analysis [9]. Using a partitioning of the total information, we observe how each process sustains its dynamics and its level of dependency on other components in the multivariate system (Figure 4.5). For instance, for the horizontal wind speed, WS, the unique information from self-dependency, $U_{self,\mathscr{T}}$, is the main contributor sustaining its dynamics across clusters, followed by the synergistic, $S_{\mathscr{T}}$, and unique information from cross-dependencies, $U_{cross,\mathscr{T}}$. However, the cross-dependencies are particularly significant for WS in cluster #3, with contributions coming from both $S_{\mathscr{T}}$ and $U_{cross,\mathscr{T}}$, in comparison to the other two clusters. In the case of CO_2 and H_2O, the redundant information, $R_{\mathscr{T}}$, is the main contributor across clusters. Lastly, marked variables as "self-dependent" refer to variables that rely only on their own dynamics. For example, the vertical wind speed, U_z, is causally a self-dependent variable regardless of the cluster. This means that the history of the variable itself, rather than the causal dependence on other variables, determines the observed present state.

Now, we explore these findings in more detail regarding their role in self-or cross-dependencies that dominate these processes, by exploring their origin in the immediate or distant histories of the variables. From our results, we can infer some unique characteristics of the system. In Figure 4.6, the analysis of the immediate $\mathscr{J}$ and distant $\mathscr{D}$ causal histories provides us with three main results. First, we notice that the processes are mainly sustained by self-dependant variables in the short-term dynamics. We find that specially for the horizontal and vertical wind components, WS and U_z, which reflect the dynamic character of the atmosphere, the unique information from self-dependency in the immediate history, $U_{self,\mathscr{J}}$, is the main contributor for sustaining its dynamics. On the other hand, the redundant information in the immediate history, $R_{\mathscr{J}}$, that refers to the overlapping information from the sources is dominant for CO_2 and H_2O.

Second, from our results we can classify the variables based on their origin, whether they are driven by short-term local or non-local conditions (Figure 4.6). Variables driven by non-local conditions refer to variables with strong self-dependency, such as U_z and T, which dynamics are primarily or even solely supported by its own causal history, without the influence of other variables. This category also includes WS, which is mainly influenced by its unique information from self-dependency in the immediate and distant history, $U_{self,\mathscr{J}}$ and $U_{self,\mathscr{D}}$, respectively. Therefore, we can infer that the variables that control the dynamic character of the atmosphere, WS and U_z, as well as the

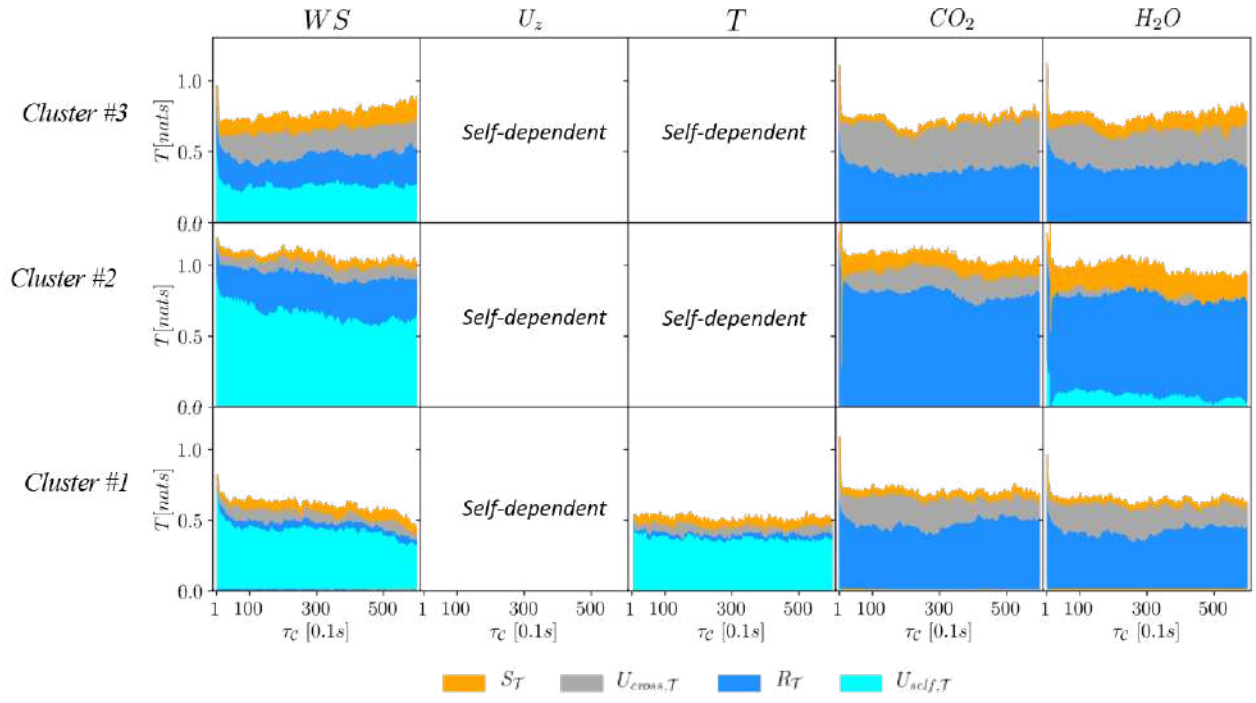

Figure 4.5: Partial information decomposition (PID) for the total causal history for the DAGs for high frequency time series as shown in Figure 4.2, for cluster #3 (top), cluster #2 (middle), and cluster #1 (bottom). The total information is decompose into the complementary synergistic information, S_T, the information only provided in the presence of all variables; the redundant information, R_T, the overlapping information from the sources; and the unique contributions from the self, $U_{self,T}$ and cross-dependencies, $U_{cross,T}$.

thermodynamics, T, are driven by non-local conditions. On the other hand, CO_2 and H_2O, which reflect the scalar transport at the biosphere-atmosphere interface, are mainly driven by short-term local conditions given that the redundant information for the immediate history dominates, $R_{\mathcal{J}}$, and therefore their dynamics are sustained by the influence of other multiple variables in their causal history in the same manner and not only by themselves.

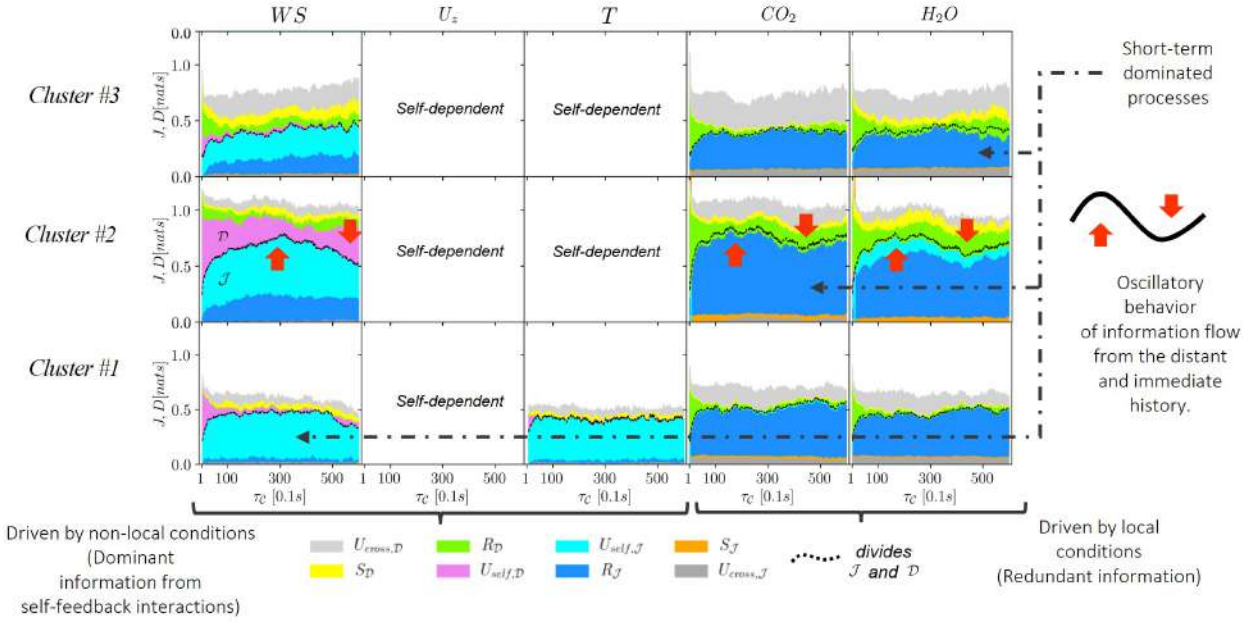

Figure 4.6: Partitioning of information using PID for the self- and cross dependencies in the immediate, $\mathscr{J}$, and distant, $\mathscr{D}$, causal histories [9], [47] of each variable (in columns) and cluster (in rows) based on the directed acyclic graphs for time series in Figure 4.2. The information is decompose into the complementary synergistic information, S_T; redundant information, R_T; and the unique contributions from self, U_{self} and cross-dependencies, U_{cross}. The black dotted lines set the division between the immediate, $\mathscr{J}$ (bellow the black-dotted line), and distant histories, $\mathscr{D}$ (above the black-dotted line), respectively.

Third, our results also show that as τ_C increases the total information is of the same order of magnitude across variables in the same cluster (Figure 4.5 and 4.6). We notice that the variables in cluster #2 are influenced by a larger amount of information through their causal histories in comparison with clusters #1 and #3. Cluster #2 shows the causal structure of the average dynamics of the cluster under the influence of an intervention on the dynamics of the whole system. Such intervention causes a temporary decrease in the connectivity of the system, which refers to less number of edges representing interdependencies among variables (Figure 4.2)[134]. Such a decrease in connectivity favors a larger amount of information flowing among components of the system (Figures 4.5 and 4.6). Our results suggest that the system potentially optimizes its causal structure by reducing its dimensionality to facilitate a larger amount of information flow. This finding suggests the existence of preferential channels for information flow: an attribute that links the causal structure and function of the system at turbulence scales. Therefore, we argue that such a minimal and optimal

arrangement of the causal structure is shaped by the system to efficiently transfer information among components of the system.

Moreover, based on our results, we argue that the system shapes its causal structure as a mechanism for resilience to preserve its functionality during an intervention. We previously discussed that the system has the ability to temporarily modify its causal structure to optimally meet its essential function, the transfer of information. We argue that the ecohydrological system at turbulence scales is resilient in the sense that it evolves adjusting its causal structure to protect its critical functionality from disruption. Here, resilience refers to the ability of the system to successfully adapt to interventions maintaining the functionality of the system. These support the concept of "system resilience to information flow", proposed by Hernandez Rodriguez and Kumar [134], which refers to the ability of a system to remain in a stable range of the organizational state of its causal structure that sustains the flow of information, even under the effect of an intervention.

Based on the findings on Hernandez Rodriguez and Kumar [134] regarding the evolution of the causal structure followed with the causal history analysis implemented here, we argue that the system copes with an intervention using two strategies. First, the system can adjust to a minimal arrangement of causal interactions to favor a larger information flow from the causal history of each variable, as previously discussed. Second, although the system is mainly sustained by short-term dominated processes, it temporarily allows a significant support from the distant history to maintain its dynamics (Figure 4.6, cluster #2). Notice that the total information among clusters is larger for WS, CO_2, and H_2O, which are the variables that support their dynamics in both the immediate and distant causal history (Figure 4.4). On the other hand, the self-dependant variables, U_z and T, highly depend on their immediate causal history regardless the cluster.

We have tested our hypothesis, finding different functional responses when comparing among clusters. We build upon our causal structure analysis [134] to postulate that in the evolutionary dynamics of the ecohydrological system at turbulence scales, functional differences in the information flow of the system are the reflection of causal structural differences promoted by the system itself to maintain its dynamics, where functional differences are detected by the amount and type of dependencies using a SUR approach in a causal history framework.

4.3.4 Causal interdependencies as the cause for long-range dependencies

In our results for cluster #2, we noticed an oscillatory behavior in the information transferred from the immediate and distant histories as a function of τ_C, for both set of variables that have local and non-local influence. This particular behavior was first noticed in the ratio between the distant and total information for temperature, a predominantly self-dependent variable (Figure 4.4). Later, we observe the same behavior for the cross-dependent variables in the PID analysis (Figure 4.6). For instance, for the horizontal wind speed, WS, the contribution of the unique information from

self-dependency in the immediate history, $U_{self,\mathscr{I}}$ shows an oscillatory negative correlation with the unique information from self-dependency in the distant history, $U_{self,\mathscr{D}}$ (Figure 4.6). This behavior is also observed for the cross-dependent local variables, CO_2 and H_2O, which share a similar periodic signal for the redundant information, $R_{\mathscr{I}}$. Therefore, in terms of information flow from the causal history, this oscillatory behavior reveals that the system can experience transitions in the ratio of the immediate to distant history over time.

Our results suggest that the oscillatory behavior in cluster #2 is related to the frozen turbulence described by Taylor's hypothesis [87]. Eddies could be changing size and shape as they drift by the sensor. Taylor's hypothesis holds when the turbulence intensity is small relative to the mean wind speed [87]. As stated in Hernandez Rodriguez and Kumar [134], in cluster #2 (Figures 3.7e and 3.7h in chapter 3), the dynamics of the horizontal WS, and vertical wind speed, U_z, are decoupled from the thermodynamics, T. In such stagnant environment in which the system temporarily exist, the mean wind is not contributing to the vertical movement of scalars, such that eddies can change in composition as are advected past the sensor, resulting in a larger amount of information coming alternatively from the immediate and distant history of each variable carried by the eddies.

Looking at the spectral behavior of the data, we observe that the high frequency data exhibit fractal $1/f$ scaling (Figure 4.7), which refers to the inverse proportionality between spectral power and frequency [8]. We also observe the prevalence of $1/f$ noise [138], also called Flicker noise [139] or pink noise spectrum, which is characterized by a wide, flat plateau at high frequencies. Typically, the effect of periodic thermal or electrical fluctuations over time are used to explain the $1/f$ scaling noise in time series [140], but no causal relationship has been previously established regarding the information flow of the system.

We argue that the self-dependency dominance in the causal history causes the absence of power-law dependence in the spectral plot of a variable. When there is no power-law dependence, we observe a noise plateau at high frequencies, such as in the case of temperature and the vertical wind velocity component, U_z (Figure 4.7). In other words, the temperature is flattened out at high frequencies in the spectral plot, and therefore, temperature is not following the turbulence power law. There is a basis to infer that when cross-dependent short-term dominated processes increase its self-dependency on its distant history, some long-range dependencies appear. For instance, in cluster #2 we observe the oscillatory behavior for all variables and power-law independence at high frequencies caused by an increase in the self-dependence on its distant history. In the case of WS, it is due to the additional unique information from self-dependency in the distant history, $U_{self,\mathscr{D}}$. For CO_2 and H_2O, is due to the larger dependence on the redundant information from the distant history, $R_{\mathscr{D}}$ (Figure 4.6). Therefore, we postulate that increments of the self-dependence of a variable, especially relying on its distant causal history, is reflected in its spectral behavior as a lack of power-law dependency.

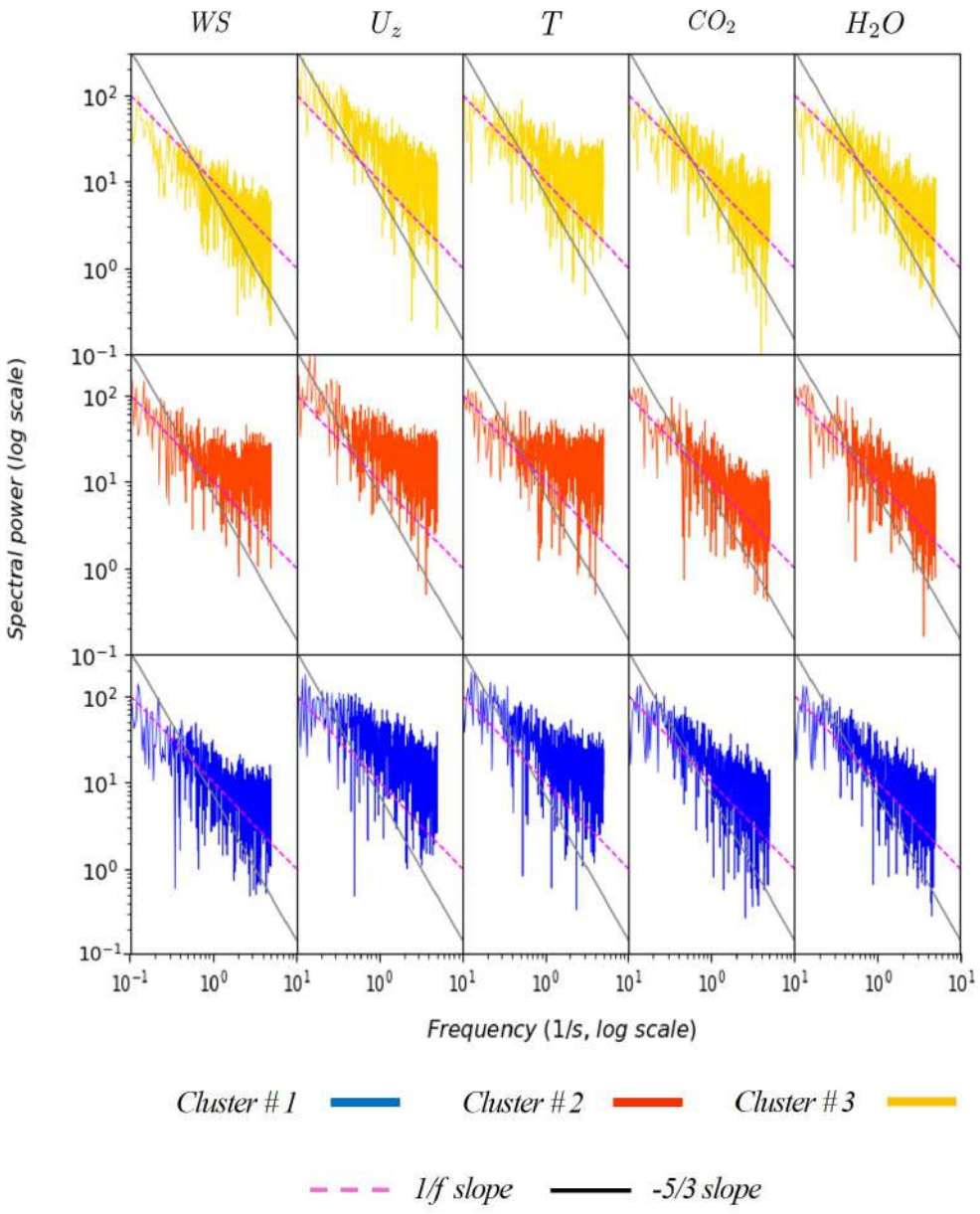

Figure 4.7: Fourier power spectrum of the time series of variables at high frequency. Each row refers to each of the clusters in Figure 4.2. The spectra show a noise plateau at high frequencies and $1/f$ slope at low frequencies (magenta dashed line). The Kolmogorov spectrum slope (solid black line), has the form of $E \propto k^{-5/3}$, which is supposed to fit most of the behavior at turbulence scale [64], [141]

.

4.4 Discussion: Duality between causal structure and functionality

Causal structure and its functionality form a relationship that evolves to meet the information flow needs of the system. When we look at the dynamics of turbulence at the biosphere-atmosphere interface, carbon dioxide, water vapor, and other scalars, they are transported away from the Earth's surface in a well-developed physical mechanism called ejection-sweep [142]. In turbulence, coherent structures called eddies exhibit recurrent patterns in the presence of a boundary, such as the Earth surface [87], [143]. Momentum and energy-containing eddies, which differ on many orders of magnitude, self-sustain the cycle of inception of new eddies and there is dissipation [65], [87]. In Lozano-Durán, Bae, and Encinar [65], causal interactions were estimated among eddies using numerical simulations to determine how the knowledge of the past states of eddies reduces the uncertainty of their future states. Their results show that energy-eddies in the buffer and logarithmic layers in the atmosphere are similar and independent of the eddy size. Here we have taken the causal look further using the dynamics of turbulence at the biosphere-atmosphere interface, beyond the univariate analysis of the exploration of the behavior of wind features, by looking at the fluctuations of a multivariate complex system that encode information about interactions in the time series at high frequency. In this multivariate system, temperature, scalars such as CO_2 and H_2O, along with wind velocity components, create a causal structure of interdependencies that sustain information flow and evolves inside a range of dynamical behavior instead of being in a fixed configuration [134]. We based our analysis on the hypothesis that causal relationships are affected through information flow, and the interaction arising due to information flow sustains the dynamics of the ecohydrological system at high frequency, its functionality. Moreover, the causal structure shapes the network of information flow among variables in the multivariate complex system and serves as the foundation for the functional representation of the evolution of the dynamics of the system.

4.5 Conclusions

We explore how the causal structure of the dynamics of an ecohydrological system is related to its functionality. We use high frequency observations and information metrics to infer causality from interactions among multiple components of an ecohydrological system. Here we see the propagating fluctuations among interacting components in the system as information flow. Using the DAGs resulting from the clustering analysis to determine the evolution of the system causal structure [134], our results show differences in their information flow. We conclude that functional differences in the information flow of the system are a reflection of causal structural differences promoted by the system itself to maintain its dynamics.

We demonstrate how information theory provides a framework to determine the causal structure and functionality of ecohydrological systems. The IT-based methods developed and/or used here allow the inference of the relationship

between the causal structure and functionality of the evolving system. Based on the results of the causal structure analysis in Hernandez Rodriguez and Kumar [134], and using a causal history approach [9], [50] that implements a partial information decomposition framework [47], [48], our results suggest that at a turbulence scale, the processes that shape the dynamics of the ecohydrological system at the biosphere-atmosphere interface are mainly sustained by short-term dynamics. In addition, each process can have a predominantly local origin, such as for CO_2 and H_2O, or nonlocal origin, in the case of WS, U_z, and T. Our results show that the causal structure can be shaped by the system to meet the information flow needs of the evolving system. In addition, we find that the system can reduce its dimensionality to facilitate the transfer of large amounts of information among its components, especially under the effect of an intervention. Ultimately, we found that the cause for long-term dependencies in turbulence that do not follow a power law in the spectral plot, are caused by self-dependencies coming from the distant causal history, even when the dynamics are mainly sustained by its short-term dynamics. In the study of causal inference of ecohydrological systems, we recognized the importance of developing and implementing causal approaches to infer the dynamics of processes when using time series and other informative datasets. In the machine learning and big data era, we recognize the need of implementing causal inference in the geosciences. We aim that the tools for causal inference developed and/or implemented here to determine the evolution of the causal structure and the relationship with the functionality of the system. These analyses reveal the dynamics of the multivariate system at high frequency and provide some new avenues for causality to be integral to artificial intelligence and machine learning-driven approaches, which are also sustained and rely on observations.

Chapter 5

Conclusions

This dissertation focuses on understanding how the information carried in the observations from an eddy covariance tower encapsulates information about changes in the structure, evolution, and functionality of the system at different scales. The specific goals of this dissertation are to (i) determine in what way the composition of the heterogeneous organization of the land cover inside the flux footprint contributes to the observations measured at the tower; (ii) develop a framework for the study of the evolution of the causal structure of multivariate interactions at the turbulence scale by exploiting the high frequency measurements, and (iii) determine the relationship between the evolution of the causal structure and its relationship to the system's functionality.

In this Chapter, we summarize the general findings and key results from the research. We also discuss some potential avenues to build upon this research.

5.1 Role of organized land cover heterogeneity inside the flux footprint

A major challenge when using flux tower data refers to the unknown origin of the sources in the ecosystem surrounding the tower. One key feature associated with the flux observations is that beyond the magnitude and temporal variability characteristic of each variable, the data also carry the influence of the source/sink behavior inside its flux footprint. This is a fundamental aspect of the observations for their correct interpretation.

In Chapter 2, using data from a tall tower in the Midwestern US, we determined the composition of the land cover inside the flux footprint and characterized the effects of this heterogeneity on the observations. We focused on the role of the "organized land cover heterogeneity", which we defined as the human-induced spatial organization of the landscape, on the fluxes exchanged between the landscape and the atmosphere when measured at a tall eddy covariance tower.

When the land cover is heterogeneous, inconsistencies in data interpretation can arise when only accounting for the vegetation type in the nearest field, or alternatively assuming that multiple crop types contribute equally to the observed flux. We combined flux footprint and ecohydrological modeling together with flux tower data, to improve upon the understanding that could be obtained given any single information source. We assigned the relative contribution of the land cover types from the flux footprint to the latent heat (LE), sensible heat (H), and CO_2 fluxes (F_c). We estimated the spatial extent contributing to every flux measurement from April 2016 to April 2019, and we combined them with an ecohydrological model to obtain the temporally varying contribution of fluxes from different land covers. Then, we estimated how they contributed to the overall flux, which depends on how the crop fields are spatially distributed as "visible" in the flux footprint. We considered a structured heterogeneity in the flux footprint whose contributions are dynamically changing at every measurement time step (i.e., 15 min), instead of using the climatology of the flux footprint. Our results help explain how the contributions from different crops to the observed fluxes at the tower vary from year to year. The key results are the following:

- We found that the tower mostly "sees" fields located between 168m and 268m away, although areas as far away as several kilometers contribute to the observations. Maize fields contributed more than soybean fields during the 2016 and 2018 growing seasons, and vice versa during 2017.
- We compared our results to a hypothetical case where all vegetation is randomly distributed on the surface. We found that the crop fields that the tower "sees" account for up to 24.5% more CO_2 flux (F_c) than if the vegetation was randomly distributed on the landscape.
- We integrated ecohydrological and flux footprint modeling efforts with eddy covariance flux tower data and high temporal-spatial resolution remote sensing data to better reflect the temporal dynamics of surface fluxes. This research is of practical interest for human-induced activities at the land-surface such as agriculture and of theoretical interest regarding the connection of soil-vegetation-atmosphere interactions and its responses in terms of energy, water, and CO_2 in the near-surface layer of the Earth's Critical Zone.

5.2 Evolution of the causal structure of interactions in high frequency turbulence

In Chapter 3, we used the high frequency time series of turbulence observations (i.e., 10Hz) from a 25m tall eddy covariance flux tower to explore the evolutionary dynamics of the system. The causal structure refers to a network representation that shows the connectivity of the information transfer among components of the system. We quantified the network of inter-dependencies using directed acyclic graphs (DAG) from multivariate high frequency time series and quantified its evolutionary dynamics. We developed a novel approach to find patterns of similar behavior using these

DAGs. We explored the lagged influences resulting from self-feedback and cross-dependencies that shape the causal structure of a turbulent system. We estimated the evolution of the causal structure of the turbulence system for a clear sky day and under the influence of the rapid but progressive solar radiation limitation and recovery experienced during a solar eclipse, a unique natural intervention on the dynamics of the whole system. Our key contributions and findings are:

- This is the first work to holistically quantify the evolution of the causal structure of a complex system at the turbulence scale.
- We moved beyond the traditional spectral and/or correlative analyses when studying turbulence to explore causal dependency among its components. Our implementation at the biosphere-atmosphere interface uses the information carried in high frequency data, allowing the study of a multivariate view of the turbulent system.
- Our novel approach found patterns of similar behavior using DAGs, which is based on a distance-based classification to characterize the structural differences between the DAGs and a k-means clustering approach that selects an unbiased number of clusters.
- We found the formation of behavioral patterns of similar causal structure. We found three well-defined clusters of structural connectivity in the system. Each cluster's centroid can be seen as representing the average dynamics within a cluster of DAGs of similar dynamics.
- From the clusters, we found that the dynamics of the horizontal wind speed are largely self-dependant and that the air temperature T strongly influences the scalar transport of CO_2 and H_2O. We also identified the distinguishable features for each of the clusters, such as for cluster #2, which is mainly characterized by a decouple between the two main attributes controlling the dynamics of the whole system at turbulent scales: the dynamic influence of the atmosphere given by the wind speed, and the influence of the thermodynamics, which has a strong causal influence on the scalar transport.
- From our observations of how the connectivity of the system may "look like" under a clear sky day and under the effect of an intervention, we observed that the variables in the turbulent system exist and inform each other in a persistent re-organizational state in a range of dynamical behavior instead of being in a fixed configuration. Therefore, the connectivity of the complex system fluctuates within a range of variations, which allows us to define patterns of evolution of its dynamics.
- Under the effects of a rapid intervention such as a solar eclipse, which is able to perturb the dynamics of the whole system, the connectivity defined by the number of links that allow information flow within the system decreases momentarily and eventually reaches a new dynamical state where the connectivity is restored.

- Our results led us to define the “system resilience to information flow", that reflects the ability of a system to continue functioning when there has been a disruption or intervention in the typical behavior of the system to ensure the continuity of information flow.

5.3 Evolution of causal structure - function coupling

In Chapter 4, information theory (IT) provides a framework to determine the coupling between the causal structure and the functionality of the evolving system. We defined functionality as the nature of the interactions that can be partitioned and classified as unique, synergistic, and redundant (SUR) information flow, through the partial information decomposition (PID) approach [47], [48] using momentary partial information decomposition (MPID) [54]. We based our analysis on the causal structure analysis from Chapter 3 [134], and we used the causal history approach [9], [50] that implements a partial information decomposition framework [47], [48] to determine if the evolution of the causal structure reflect different underlying dynamics. Our key findings are:

- Our results suggest that at a turbulent scale, the processes that shape the dynamics of the ecohydrological system at the biosphere-atmosphere interface are mainly sustained by short-term dynamics.

- We classified the variables in the system based on their origin, whether they are driven by short-term local or non-local conditions. Variables driven by non-local conditions refer to variables with strong self-dependency, whose dynamics are primarily or even solely supported by their own causal history, without the influence of other variables. We found that the variables that control the dynamic character of the atmosphere, WS and U_z, as well as the thermodynamics, T, are driven by non-local conditions. On the other hand, the variables driven by short-term local conditions are highly supported by the influence of other variables and not only by themselves, such as for CO_2 and H_2O, which reflect the scalar transport at the biosphere-atmosphere interface.

- From comparing the connectivity in the causal structure and the associated amount of information flow in each of the clusters centroids, we postulated that the causal structure can be shaped by the system to meet the information flow needs of the evolving system. We found that the system can reduce its dimensionality to facilitate the transfer of large amounts of information among its components, especially under the effect of an intervention.

- Ultimately, we found that the cause for long-term dependencies in turbulent systems that do not follow a power law in the spectral plot, are caused by self-dependencies coming from the distant causal history, even when the dynamics are mainly sustained by its short-term dynamics.

5.4 Avenues for future research

This dissertation presented how the information carried in the observations about the composition of the land cover is relevant to the correct interpretation of the flux observations obtained at a eddy-covariance flux tower. Moreover, this dissertation explored the information contained in the flux tower observations to obtain a deeper understanding of the complex system dynamics when using the information carried by the observations. Our novel approach for the study of the evolution of the causal structure of multivariate interactions, and the relationship between the evolution of the causal structure and the system's functionality is the first contribution of its type and can be implemented to other complex systems.

Given the worldwide concern of the rapid change of the earth's atmosphere and biosphere since the industrial revolution, the study of the behavior of the CO_2 concentration, along with other gasses, continues to be an important aspect of research. The increasing availability of observations of systems around the world, offer us an opportunity to create new and innovative avenues to explore systems of different levels of complexity. All these efforts are of significant relevance to better understand the biosphere-atmosphere interface dynamics from the observations we collect worldwide. This is especially urgent given the many interventions and dramatic changes that are currently experienced in all ecosystems. Unless we investigate the origin and dynamics of the signals we measure from natural and agricultural systems, we will not be able to effectively account for actions towards improving CO_2 sequestration, and other techniques to help reduce gases in the biosphere-atmosphere interface.

In the era of data, and artificial intelligence and machine learning, theoretical and experimental efforts must be made towards considering cause-effect relationships, to find the reasons behind the predictions that are ultimately used for decision making. The clustering approach for DAGs for time series developed here, which is based on information theoretical metrics, contributes to paving the way for near-term research that will allow a next level of understanding of complex systems. Inspirational works based on information theory and complex systems, such as the ones recently awarded with the Nobel Prize in Physics 2021 "for groundbreaking contributions to our understanding of complex systems"[1] to Syukuro Manabe and Klaus Hasselmann "for the physical modelling of Earth's climate, quantifying variability and reliably predicting global warming" and to Giorgio Parisi "for the discovery of the interplay of disorder and fluctuations in physical systems from atomic to planetary scales", will continue encouraging more researchers for the development of causal models. Studies such as "machine reconstruction" [108], a set of transition rules built from observational data denominated as "epsilon machine" [144], [145], produces a finite-state machine representation directly from observational data and displays the computational structure of the process being modeled. We anticipate that in the near-future, the integration of physically- and data-based techniques with causal models will contribute to the further understanding of complex systems such as turbulence, under observed conditions and allowing the use of

[1] https://www.nobelprize.org/prizes/physics/2021/summary/

counterfactuals [57] to explore the feasible but not yet observed behavior of the system, to better determine its dynamics and evolution in real world settings with high complexity.

References

[1] D. D. Baldocchi, "Assessing the eddy covariance technique for evaluating carbon dioxide exchange rates of ecosystems: Past, present and future," *Global change biology*, vol. 9, no. 4, pp. 479–492, 2003.

[2] M. Jung, M. Reichstein, H. A. Margolis, *et al.*, "Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations," *Journal of Geophysical Research: Biogeosciences*, vol. 116, no. G3, 2011.

[3] M. Aubinet, T. Vesala, and D. Papale, *Eddy covariance: A practical guide to measurement and data analysis*. Springer Science & Business Media, 2012.

[4] D. Baldocchi, E. Falge, L. Gu, *et al.*, "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*, vol. 82, no. 11, pp. 2415–2434, 2001.

[5] D. Schimel, W. Hargrove, F. Hoffman, and J. MacMahon, "Neon: A hierarchically designed national ecological network," *Frontiers in Ecology and the Environment*, vol. 5, no. 2, pp. 59–59, 2007.

[6] G. Burba, *Eddy covariance method for scientific, industrial, agricultural and regulatory applications: A field book on measuring ecosystem gas exchange and areal emission rates*. LI-Cor Biosciences, 2013.

[7] G. Tononi and G. M. Edelman, "Consciousness and complexity," *science*, vol. 282, no. 5395, pp. 1846–1851, 1998.

[8] J. W. Kirchner and C. Neal, "Universal fractal scaling in stream chemistry and its implications for solute transport and water quality trend detection," *Proceedings of the National Academy of Sciences*, vol. 110, no. 30, pp. 12 213–12 218, 2013.

[9] P. Jiang and P. Kumar, "Using information flow for whole system understanding from component dynamics," *Water Resources Research*, 2019.

[10] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman, *Causation, prediction, and search*. MIT press, 2000.

[11] H. Chu, X. Luo, Z. Ouyang, *et al.*, "Representativeness of eddy-covariance flux footprints for areas surrounding ameriflux sites," *Agricultural and Forest Meteorology*, vol. 301, p. 108 350, 2021.

[12] D. Schimel, F. D. Schneider, J. Carbon, and E. Participants, "Flux towers in the sky: Global ecology from space," *New Phytologist*, vol. 224, no. 2, pp. 570–584, 2019.

[13] J.-P. Tuovinen, M. Aurela, J. Hatakka, *et al.*, "Interpreting eddy covariance data from heterogeneous siberian tundra: Land-cover-specific methane fluxes and spatial representativeness," *Biogeosciences*, vol. 16, no. 2, pp. 255–274, 2019.

[14] M. Y. Leclerc and T. Foken, *Footprints in micrometeorology and ecology*. Springer, 2014, vol. 239.

[15] F. Pasquill, "Some aspects of boundary layer description," *Quarterly Journal of the Royal Meteorological Society*, vol. 98, no. 417, pp. 469–494, 1972.

[16] T. Vesala, N. Kljun, Ü. Rannik, *et al.*, "Flux and concentration footprint modelling: State of the art," *Environmental Pollution*, vol. 152, no. 3, pp. 653–666, 2008.

[17] P. Schuepp, M. Leclerc, J. MacPherson, and R. Desjardins, "Footprint prediction of scalar fluxes from analytical solutions of the diffusion equation," *Boundary-Layer Meteorology*, vol. 50, no. 1-4, pp. 355–373, 1990.

[18] N. Kljun, P. Calanca, M. Rotach, and H. P. Schmid, "A simple two-dimensional parameterisation for flux footprint prediction (ffp)," *Geoscientific Model Development*, vol. 8, no. 11, pp. 3695–3713, 2015.

[19] A. Griebel, D. Metzen, E. Pendall, G. Burba, and S. Metzger, "Generating spatially robust carbon budgets from flux tower observations," *Geophysical Research Letters*, vol. 47, no. 3, e2019GL085942, 2020.

[20] S. Fan, S. C. Wofsy, P. Bakwin, *et al.*, "Micrometeorological measurements of ch4 and co2 exchange between the atmosphere and subarctic tundra," *Journal of Geophysical Research: Atmospheres*, vol. 97, no. D15, pp. 16 627–16 643, 1992.

[21] I. Forbrich, L. Kutzbach, C. Wille, T. Becker, J. Wu, and M. Wilmking, "Cross-evaluation of measurements of peatland methane emissions on microform and ecosystem scales using high-resolution landcover classification and source weight modelling," *Agricultural and forest meteorology*, vol. 151, no. 7, pp. 864–874, 2011.

[22] J. Chi, M. B. Nilsson, H. Laudon, *et al.*, "The net landscape carbon balance—integrating terrestrial and aquatic carbon fluxes in a managed boreal forest landscape in sweden," *Global change biology*, vol. 26, no. 4, pp. 2353–2367, 2020.

[23] J. Chi, M. B. Nilsson, N. Kljun, *et al.*, "The carbon balance of a managed boreal landscape measured from a tall tower in northern sweden," *Agricultural and Forest Meteorology*, vol. 274, pp. 29–41, 2019.

[24] T. Biermann, W. Babel, W. Ma, *et al.*, "Turbulent flux observations and modelling over a shallow lake and a wet grassland in the nam co basin, tibetan plateau," *Theoretical and applied climatology*, vol. 116, no. 1, pp. 301–316, 2014.

[25] S. Kirby, R. Dobosy, D. Williamson, and E. Dumas, "An aircraft-based data analysis method for discerning individual fluxes in a heterogeneous agricultural landscape," *agricultural and forest meteorology*, vol. 148, no. 3, pp. 481–489, 2008.

[26] L. C. Hernandez Rodriguez, A. E. Goodwell, and P. Kumar, "Inside the flux footprint: Understanding the role of organized land cover heterogeneity on land-atmospheric fluxes," *This chapter is in review for Agricultural and Forest Meteorology*, 2022.

[27] C. G. Wilson, B. Abban, L. L. Keefer, *et al.*, "The Intensively Managed Landscape Critical Zone Observatory: A scientific testbed for understanding critical zone processes in agroecosystems," *Vadose Zone Journal*, vol. 17, no. 1, 2018.

[28] P. V. Le and P. Kumar, "Power law scaling of topographic depressions and their hydrologic connectivity," *Geophysical Research Letters*, vol. 41, no. 5, pp. 1553–1559, 2014.

[29] E. Foote, "Art. xxxi.–circumstances affecting the heat of the sun's rays," *American Journal of Science and Arts (1820-1879)*, vol. 22, no. 66, p. 382, 1856.

[30] J. D. Ortiz and R. Jackson, "Understanding eunice foote's 1856 experiments: Heat absorption by atmospheric gases," *Notes and Records*, 2020.

[31] J. London, *A study of the atmospheric heat balance*, 287. Research Division, College of Engineering, Department of Meteorology and . . ., 1962, vol. 57.

[32] W. Wang, Y. Yung, A. Lacis, T. Mo, and J. Hansen, "Greenhouse effects due to man-made perturbations of trace gases," *Science*, vol. 194, no. 4266, pp. 685–690, 1976.

[33] J. F. Mitchell, "The "greenhouse" effect and climate change," *Reviews of Geophysics*, vol. 27, no. 1, pp. 115–139, 1989.

[34] T. R. Anderson, E. Hawkins, and P. D. Jones, "Co2, the greenhouse effect and global warming: From the pioneering work of arrhenius and callendar to today's earth system models," *Endeavour*, vol. 40, no. 3, pp. 178–187, 2016.

[35] D. W. Kweku, O. Bismark, A. Maxwell, *et al.*, "Greenhouse effect: Greenhouse gases and their impact on global warming," *Journal of Scientific research and reports*, pp. 1–9, 2017.

[36] P. H. Gleick, "Climate change, hydrology, and water resources," *Reviews of Geophysics*, vol. 27, no. 3, pp. 329–344, 1989.

[37] S. Bastianoni, F. M. Pulselli, and E. Tiezzi, "The problem of assigning responsibility for greenhouse gas emissions," *Ecological Economics*, vol. 49, no. 3, pp. 253–257, 2004, ISSN: 0921-8009. DOI: https://doi.org/10.1016/j.ecolecon.2004.01.018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0921800904001090.

[38] J. M.-F. Johnson, A. J. Franzluebbers, S. L. Weyers, and D. C. Reicosky, "Agricultural opportunities to mitigate greenhouse gas emissions," *Environmental Pollution*, vol. 150, no. 1, pp. 107–124, 2007, ISSN: 0269-7491. DOI: https://doi.org/10.1016/j.envpol.2007.06.030. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0269749107003016.

[39] M. Prokopenko, F. Boschetti, and A. J. Ryan, "An information-theoretic primer on complexity, self-organization, and emergence," *Complexity*, vol. 15, no. 1, pp. 11–28, 2009. DOI: https://doi.org/10.1002/cplx.20249. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cplx.20249. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cplx.20249.

[40] P. Kumar and B. L. Ruddell, "Information driven ecohydrologic self-organization," *Entropy*, vol. 12, no. 10, pp. 2085–2096, 2010.

[41] P. Kumar, "Hydrocomplexity: Addressing water security and emergent environmental risks," *Water Resources Research*, vol. 51, no. 7, pp. 5827–5838, 2015.

[42] T. Schreiber, "Measuring information transfer," *Physical review letters*, vol. 85, no. 2, p. 461, 2000.

[43] J. Runge, J. Heitzig, V. Petoukhov, and J. Kurths, "Escaping the curse of dimensionality in estimating multivariate transfer entropy," *Physical review letters*, vol. 108, no. 25, p. 258 701, 2012a.

[44] B. L. Ruddell and P. Kumar, "Ecohydrologic process networks: 1. identification," *Water Resources Research*, vol. 45, no. 3, 2009.

[45] A. Goodwell and P. Kumar, "Information theoretic measures to infer feedback dynamics in coupled logistic networks," *Entropy*, vol. 17, no. 11, pp. 7468–7492, 2015.

[46] B. L. Ruddell and P. Kumar, "Ecohydrologic process networks: 2. analysis and characterization," *Water Resources Research*, vol. 45, no. 3, 2009.

[47] A. E. Goodwell and P. Kumar, "Temporal information partitioning networks (tipnets): A process network approach to infer ecohydrologic shifts," *Water Resources Research*, vol. 53, no. 7, pp. 5899–5919, 2017.

[48] P. L. Williams and R. D. Beer, "Nonnegative decomposition of multivariate information," *arXiv preprint arXiv:1004.2515*, 2010.

[49] A. E. Goodwell, P. Jiang, B. L. Ruddell, and P. Kumar, "Does information theory provide a new paradigm for earth science? causality, interaction, and feedback," *Water Resources Research*, e2019WR024940,

[50] P. Jiang and P. Kumar, "Information transfer from causal history in complex system dynamics," *Physical Review E*, vol. 99, no. 1, p. 012 306, 2019.

[51] J. Runge, "Quantifying information transfer and mediation along causal pathways in complex systems," *Physical Review E*, vol. 92, no. 6, p. 062 829, 2015.

[52] J. Runge, V. Petoukhov, J. F. Donges, *et al.*, "Identifying causal gateways and mediators in complex spatio-temporal systems," *Nature communications*, vol. 6, no. 1, pp. 1–10, 2015.

[53] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, "Detecting causal associations in large nonlinear time series datasets," *arXiv preprint arXiv:1702.07007*, 2017.

[54] P. Jiang and P. Kumar, "Interactions of information transfer along separable causal paths," *Physical Review E*, vol. 97, no. 4, p. 042 310, 2018.

[55] A. E. Goodwell, P. Kumar, A. W. Fellows, and G. N. Flerchinger, "Dynamic process connectivity explains ecohydrologic responses to rainfall pulses and drought," *Proceedings of the National Academy of Sciences*, vol. 115, no. 37, E8604–E8613, 2018.

[56] P. Jiang, "Causal history analysis of complex system dynamics," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2019.

[57] J. Pearl, *Causality*. Cambridge university press, 2009.

[58] J. Faye, "Backward Causation," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed., Spring 2021, Metaphysics Research Lab, Stanford University, 2021.

[59] P. E. Anderson, H. J. Jensen, L. Oliveira, and P. Sibani, "Evolution in complex systems," *Complexity*, vol. 10, no. 1, pp. 49–56, 2004.

[60] R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman lectures on physics, Vol. I: The new millennium edition: mainly mechanics, radiation, and heat*. Basic books, 2011, vol. 1.

[61] I. Eames and J.-B. Flor, *New developments in understanding interfacial processes in turbulent flows*, 2011.

[62] U. Frisch and A. N. Kolmogorov, *Turbulence: the legacy of AN Kolmogorov*. Cambridge university press, 1995.

[63] O. Reynolds, "Xxix. an experimental investigation of the circumstances which determine whether the motion of water shall be direct or sinuous, and of the law of resistance in parallel channels," *Philosophical Transactions of the Royal society of London*, no. 174, pp. 935–982, 1883.

[64] D. G. Ortiz-Suslow and Q. Wang, "An evaluation of kolmogorov's- 5/3 power law observed within the turbulent airflow above the ocean," *Geophysical Research Letters*, vol. 46, no. 24, pp. 14 901–14 911, 2019.

[65] A. Lozano-Durán, H. J. Bae, and M. P. Encinar, "Causality of energy-containing eddies in wall turbulence," *Journal of Fluid Mechanics*, vol. 882, 2020.

[66] C. G. Belinchon, "Multiscale information transfer in turbulence," Ph.D. dissertation, Université de Lyon, 2018.

[67] R. Cerbus and W. Goldburg, "Information content of turbulence," *Physical Review E*, vol. 88, no. 5, p. 053 012, 2013.

[68] G. Tissot, A. Lozano-Durán, L. Cordier, J. Jiménez, and B. R. Noack, "Granger causality in wall-bounded turbulence," in *Journal of Physics: Conference Series*, IOP Publishing, vol. 506, 2014, p. 012 006.

[69] P. Kumar, P. V. Le, A. T. Papanicolaou, *et al.*, "Critical transition in critical zone of Intensively Managed Landscapes," *Anthropocene*, vol. 22, pp. 10–19, 2018.

[70] M. Richardson and P. Kumar, "Critical zone services as environmental assessment criteria in intensively managed landscapes," *Earth's Future*, vol. 5, no. 6, pp. 617–632, 2017.

[71] J. Hatfield, G. Takle, R. Grotjahn, *et al.*, "Climate change impacts in the United States: The third national climate assessment," *Washington, DC*, pp. 150–174, 2014.

[72] U. NASS, "Field crops: Usual planting and harvesting dates," *USDA National Agricultural Statistics Service, Agriculural Handbook*, vol. 628, 2010.

[73] E. Foufoula-Georgiou, Z. Takbiri, J. A. Czuba, and J. Schwenk, "The change of nature and the nature of change in agricultural landscapes: Hydrologic regime shifts modulate ecological transitions," *Water Resources Research*, vol. 51, no. 8, pp. 6649–6671, 2015.

[74] Licor. "Tovi." (2021), [Online]. Available: https://licor.com/tovi/ (visited on 08/13/2021).

[75] V. Masson-Delmotte, Z. P., P. A., *et al.*, *IPCC, 2021: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 2021.

[76] A. M. Anders, E. A. Bettis III, D. A. Grimley, A. J. Stumpf, and P. Kumar, "Impacts of quaternary history on critical zone structure and processes: Examples and a conceptual model from the Intensively Managed Landscapes Critical Zone Observatory," *Frontiers in Earth Science*, vol. 6, p. 24, 2018.

[77] V. Mishra and K. A. Cherkauer, "Retrospective droughts in the crop growing season: Implications to corn and soybean yield in the Midwestern United States," *Agricultural and Forest Meteorology*, vol. 150, no. 7-8, pp. 1030–1045, 2010.

[78] D. Drewry, P. Kumar, S. Long, C. Bernacchi, X.-Z. Liang, and M. Sivapalan, "Ecohydrological responses of dense canopies to environmental variability: 1. Interplay between vertical structure and photosynthetic pathway," *Journal of Geophysical Research: Biogeosciences*, vol. 115, no. G4, 2010.

[79] R. E. Warner, S. P. Havera, L. M. David, and R. J. Siemers, "Seasonal abundance of waste corn and soybeans in Illinois," *The Journal of wildlife management*, pp. 142–148, 1989.

[80] C. Monfreda, N. Ramankutty, and J. A. Foley, "Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000," *Global biogeochemical cycles*, vol. 22, no. 1, 2008.

[81] U. NASS, "Illinois corn and soybean county estimates," 2016-2018.

[82] D. Vogelezang and A. A. Holtslag, "Evaluation and model impacts of alternative boundary-layer height formulations," *Boundary-Layer Meteorology*, vol. 81, no. 3, pp. 245–269, 1996.

[83] T. R. Lee and S. F. De Wekker, "Estimating daytime planetary boundary layer heights over a valley from rawinsonde observations at a nearby airport: An application to the page valley in virginia, united states," *Journal of Applied Meteorology and Climatology*, vol. 55, no. 3, pp. 791–809, 2016.

[84] D. J. Seidel, Y. Zhang, A. Beljaars, J.-C. Golaz, A. R. Jacobson, and B. Medeiros, "Climatology of the planetary boundary layer over the continental united states and europe," *Journal of Geophysical Research: Atmospheres*, vol. 117, no. D17, 2012.

[85] ECMWF. "Era5 hourly data on single levels from 1979 to present. fifth generation of ecmwf atmospheric reanalyses of the global climate." (2018), [Online]. Available: https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=form (visited on 01/04/2021).

[86] T. Foken, "50 years of the monin–obukhov similarity theory," *Boundary-Layer Meteorology*, vol. 119, no. 3, pp. 431–447, 2006.

[87] R. B. Stull, *An introduction to boundary layer meteorology*. Springer Science & Business Media, 2012, vol. 13.

[88] A. F. Jacobs and J. H. Van Boxel, "Changes of the displacement height and roughness length of maize during a growing season," *Agricultural and Forest Meteorology*, vol. 42, no. 1, pp. 53–62, 1988.

[89] S. Gao, Z. Niu, N. Huang, and X. Hou, "Estimating the leaf area index, height and biomass of maize using hj-1 and radarsat-2," *International Journal of Applied Earth Observation and Geoinformation*, vol. 24, pp. 1–8, 2013.

[90] P. Alekseychik, A. Korrensalo, I. Mammarella, T. Vesala, and E.-S. Tuittila, "Relationship between aerodynamic roughness length and bulk sedge leaf area index in a mixed-species boreal mire complex," *Geophysical Research Letters*, vol. 44, no. 11, pp. 5836–5843, 2017.

[91] F. Pasquill and F. Smith, *Atmospheric Diffusion, third ed*. Wiley, New York, 1983.

[92] T. Horst and J. Weil, "Footprint estimation for scalar flux measurements in the atmospheric surface layer," *Boundary-Layer Meteorology*, vol. 59, no. 3, pp. 279–296, 1992.

[93] H. Schmid, "Source areas for scalars and scalar fluxes," *Boundary-Layer Meteorology*, vol. 67, no. 3, pp. 293–318, 1994.

[94] P. V. Le, P. Kumar, D. T. Drewry, and J. C. Quijano, "A graphical user interface for numerical modeling of acclimation responses of vegetation to climate change," *Computers & Geosciences*, vol. 49, pp. 91–101, 2012.

[95] R. Myneni, Y. Knyazikhin, and T. Park, "Mod15a2h modis/terra leaf area index/fpar 8-day l4 global 500 m sin grid v006," *NASA EOSDIS Land Processes DAAC*, 2015.

[96] A. Nguy-Robertson, A. Gitelson, Y. Peng, A. Viña, T. Arkebauer, and D. Rundquist, "Green leaf area index estimation in maize and soybean: Combining vegetation indices to achieve maximal sensitivity," *Agronomy Journal*, vol. 104, no. 5, pp. 1336–1347, 2012.

[97] R. Nemani, M. White, P. Thornton, *et al.*, "Recent trends in hydrologic balance have enhanced the terrestrial carbon sink in the United States," *Geophysical Research Letters*, vol. 29, no. 10, pp. 106–1, 2002.

[98] Q. Yan, P. V. Le, D. K. Woo, T. Hou, T. Filley, and P. Kumar, "Three-dimensional modeling of the coevolution of landscape and soil organic carbon," *Water Resources Research*, vol. 55, no. 2, pp. 1218–1241, 2019.

[99] N. Lei and J. Han, "Effect of precipitation on respiration of different reconstructed soils," *Scientific reports*, vol. 10, no. 1, pp. 1–10, 2020.

[100] H. Birch, "The effect of soil drying on humus decomposition and nitrogen availability," *Plant and soil*, vol. 10, no. 1, pp. 9–31, 1958.

[101] H. Zhang and X. Wen, "Flux footprint climatology estimated by three analytical models over a subtropical coniferous plantation in southeast china," *Journal of Meteorological Research*, vol. 29, no. 4, pp. 654–666, 2015.

[102] W. J. Murphy, "Tables for weights and measurements–crops," 1981.

[103] J. C. Kaimal and J. J. Finnigan, *Atmospheric boundary layer flows: their structure and measurement*. Oxford university press, 1994.

[104] S. Adarsh and M. J. Reddy, *Multi-scale Spectral Analysis in Hydrology: From Theory to Practice*. CRC Press, 2021.

[105] D. Vickers and L. Mahrt, "The cospectral gap and turbulent flux calculations," *Journal of atmospheric and oceanic technology*, vol. 20, no. 5, pp. 660–672, 2003.

[106] L. H. Koopmans, *The spectral analysis of time series*. Elsevier, 1995.

[107] G. Padmanabhan and A. R. Rao, "Maximum entropy spectral analysis of hydrologic data," *Water Resources Research*, vol. 24, no. 9, pp. 1519–1533, 1988.

[108] A. J. Palmer, C. W. Fairall, and W. Brewer, "Complexity in the atmosphere," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, no. 4, pp. 2056–2063, 2000.

[109] C. E. Shannon and W. Weaver, "The mathematical theory of communication. university of illinois," *Urbana*, vol. 117, 1949.

[110] C. Krich, J. Runge, D. G. Miralles, *et al.*, "Estimating causal networks in biosphere–atmosphere interaction with the pcmci approach," *Biogeosciences*, vol. 17, no. 4, pp. 1033–1061, 2020.

[111] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, "Detecting and quantifying causal associations in large nonlinear time series datasets," *Science Advances*, vol. 5, no. 11, eaau4996, 2019.

[112] P. Jiang and P. Kumar, "Bundled causal history interaction," *Entropy*, vol. 22, no. 3, p. 360, 2020.

[113] A. E. Goodwell and P. Kumar, "A changing climatology of precipitation persistence across the United States using information-based measures.," *Journal of Hydrometeorology*, vol. 20, no. 8, 2019.

[114] E. H. Van Nes, M. Scheffer, V. Brovkin, *et al.*, "Causal feedbacks in climate change," *Nature Climate Change*, vol. 5, no. 5, pp. 445–448, 2015.

[115] E. Malmi, N. Tatti, and A. Gionis, "Beyond rankings: Comparing directed acyclic graphs," *Data mining and knowledge discovery*, vol. 29, no. 5, pp. 1233–1257, 2015.

[116] N. Wiener, *I am a mathematician, the later life of a prodigy: An autobiographical account of the mature years and career of Norbert Wiener and a continuation of the account of his childhood in ex-prodigy*. Doubleday, 1956, vol. 20.

[117] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: journal of the Econometric Society*, pp. 424–438, 1969.

[118] J. Runge, J. Heitzig, N. Marwan, and J. Kurths, "Quantifying causal coupling strength: A lag-specific measure for multivariate time series related to transfer entropy," *Physical Review E*, vol. 86, no. 6, p. 061 121, 2012b.

[119] J. Runge, "Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information," in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2018, pp. 938–947.

[120] L. Kozachenko and N. N. Leonenko, "Sample estimate of the entropy of a random vector," *Problemy Peredachi Informatsii*, vol. 23, no. 2, pp. 9–16, 1987.

[121] S. Frenzel and B. Pompe, "Partial mutual information for coupling analysis of multivariate time series," *Physical review letters*, vol. 99, no. 20, p. 204 101, 2007.

[122] F. Rosas, P. A. Mediano, M. Ugarte, and H. J. Jensen, "An information-theoretic approach to self-organisation: Emergence of complex interdependencies in coupled dynamical systems," *Entropy*, vol. 20, no. 10, p. 793, 2018.

[123] S. L. Lauritzen, A. P. Dawid, B. N. Larsen, and H.-G. Leimer, “Independence properties of directed markov fields,” *Networks*, vol. 20, no. 5, pp. 491–505, 1990.

[124] A. K. Jain, “Data clustering: 50 years beyond k-means,” *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[125] R. T. Aldahdooh and W. M. Ashour, “Dimk-means" distance-based initialization method for k-means clustering algorithm",” *International Journal of Intelligent Systems and Applications*, vol. 5, no. 2, 2013.

[126] M. Matsumoto and T. Nishimura, “Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator,” *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, vol. 8, no. 1, pp. 3–30, 1998.

[127] T. M. Kodinariya and P. R. Makwana, “Review on determining number of cluster in k-means clustering,” *International Journal*, vol. 1, no. 6, pp. 90–95, 2013.

[128] J. C. Dunn, “Well-separated clusters and optimal fuzzy partitions,” *Journal of cybernetics*, vol. 4, no. 1, pp. 95–104, 1974.

[129] A. Kraskov, H. Stögbauer, and P. Grassberger, “Estimating mutual information,” *Physical review E*, vol. 69, no. 6, p. 066 138, 2004.

[130] A. A. Hagberg, D. A. Schult, and P. J. Swart, “Exploring network structure, dynamics, and function using networkx,” in *Proceedings of the 7th Python in Science Conference*, G. Varoquaux, T. Vaught, and J. Millman, Eds., Pasadena, CA USA, 2008, pp. 11–15.

[131] J. P. Crutchfield, “Between order and chaos,” *Nature Physics*, vol. 8, no. 1, pp. 17–24, 2012.

[132] C. Marletto, “The science of can and can’t,” *New Scientist*, vol. 250, no. 3330, pp. 34–37, 2021.

[133] E. S. Abbot, “The causal relations between structure and function in biology,” *The American Journal of Psychology*, vol. 27, no. 2, pp. 245–250, 1916.

[134] L. C. Hernandez Rodriguez and P. Kumar, “Causal structure in high frequency turbulence at the biosphere-atmosphere interface,” *This chapter will be send to Journal of Geophysical Research*, vol. 51, no. 7, pp. 5827–5838, 2021.

[135] D. Beverly, “Phenotypic and topographical controls of ecohydrological processes from leaf to ecosystem,” Ph.D. dissertation, University of Wyoming, 2021.

[136] S. L. Gray and R. G. Harrison, “Eclipse-induced wind changes over the british isles on the 20 march 2015,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2077, p. 20 150 224, 2016.

[137] T. Foken, B. Wichura, O. Klemm, J. Gerchau, M. Winterhalter, and T. Weidinger, "Micrometeorological measurements during the total solar eclipse of august 11, 1999," *Meteorologische zeitschrift*, vol. 10, no. 3, pp. 171–178, 2001.

[138] A. Van der Ziel, "Noise in measurements," *New York: Wiley*, 1976.

[139] D. Halford, "A general mechanical model for| f| α spectral density random noise with special reference to flicker noise 1/| f|," *Proceedings of the IEEE*, vol. 56, no. 3, pp. 251–258, 1968.

[140] A. Mennella, C. Baccigalupi, A. Balbi, *et al.*, "Imaging the first light: Experimental challenges and future perspectives in the observation of the cosmic microwave background anisotropy," *arXiv preprint astro-ph/0402528*, 2004.

[141] B. Ryden, *Chapter 7 basic turbulence - ohio state university*. [Online]. Available: http://www.astronomy.ohio-state.edu/~ryden/ast825/ch7.pdf.

[142] G. Katul, G. Kuhn, J. Schieldge, and C.-I. Hsieh, "The ejection-sweep character of scalar fluxes in the unstable surface layer," *Boundary-Layer Meteorology*, vol. 83, no. 1, pp. 1–26, 1997.

[143] L. F. Richardson, *Weather prediction by numerical process*. Cambridge university press, 2007.

[144] J. P. Crutchfield, "The calculi of emergence: Computation, dynamics and induction," *Physica D: Nonlinear Phenomena*, vol. 75, no. 1-3, pp. 11–54, 1994.

[145] J. P. Crutchfield and K. Young, "Inferring statistical complexity," *Physical review letters*, vol. 63, no. 2, p. 105, 1989.

[146] G. Burba, "Illustration of flux footprint estimates affected by measurement height, surface roughness and thermal stability," *Automated weather stations for applications in agriculture and water resources management: Current use and future perspectives. World Meteorological Organization publication*, no. 1074, pp. 77–87, 2001.

[147] J. Gash, "A note on estimating the effect of a limited fetch on micrometeorological evaporation measurements," *Boundary-Layer Meteorology*, vol. 35, no. 4, pp. 409–413, 1986.

[148] N. Kljun, M. Rotach, and H. Schmid, "A three-dimensional backward lagrangian footprint model for a wide range of boundary-layer stratifications," *Boundary-Layer Meteorology*, vol. 103, no. 2, pp. 205–226, 2002.

[149] R. Kooper, B. Angelo, L. Marini, P. Kumar, and M. Muste, "Information architecture used to manage multi-domain data analysis in Intensively Managed Landscape - Critical Zone Observatory," in *American Geophysical Union, Fall Meeting 2016, abstract IN51C-1867*, 2016.

[150] D. T. Drewry, P. Kumar, S. Long, C. Bernacchi, X.-Z. Liang, and M. Sivapalan, "Ecohydrological responses of dense canopies to environmental variability: 2. role of acclimation under elevated CO_2," *Journal of Geophysical Research: Biogeosciences*, vol. 115, no. G4, 2010.

Appendix A

Supplementary information

A.1 Using a one-dimensional flux footprint approach when estimating the role of the organized land cover heterogeneity

Here we focus on the implementation of a one-dimensional flux footprint approach when estimating the role of the organized land cover heterogeneity on the fluxes exchanged between the biosphere and the atmosphere when measured by a tall eddy covariance tower. Using a one-dimensional method that can be considered simpler with respect to the more sophisticated two-dimensional approach for the flux footprint described in Chapter 2, here we study the effect of the organized land cover heterogeneity when accounting for the relative contribution of fluxes from the mosaic of different crops along the tower fetch. In this context, the fetch refers to the distance from the tower when describing the flux footprint [6], [14]. When estimating a one-dimensional flux footprint the associated fetch will shift direction at every time step, as the fetch points towards the prevalent upwind direction. The resolution of the flux tower data is 15-min. In this appendix, we explain how a simple yet meaningful one-dimensional flux footprint method can support, with some caveats, the estimation of the relative contribution of fluxes measured at the tower for each crop type along the tower fetch in a landscape characterized by an organized land cover heterogeneity. We also provide a comparison of the results of the relative contribution of the fluxes when using one-and two-dimensional flux footprints models. Finally, we conclude with the advantages and disadvantages of each method towards the estimation of the relative contribution of each crop type.

A.1.1 Estimation of the one-dimensional flux footprint

In the literature, there are several theoretical approaches available when selecting a footprint function, including: analytical model, Lagrangian stochastic particle dispersion model, large-eddy simulation, and ensemble-averaged closure model [14], [16]. Here, we selected the analytical model proposed by Schuepp, Leclerc, MacPherson, *et al.* [17], a simple one-dimensional model derived from the analytical solution of a diffusion equation for near-neutral atmospheric conditions [14], [17]. In this model, the tower fetch is located on the horizontal axis, x, while the tower height is on the vertical direction, z (Figure 2.3). The cumulative normalized contribution of flux, Ω(x), is expressed as:

$$\Omega(x) = -\int_0^x \frac{U(z-d)}{u_* k x^2} \cdot \exp\left(-\frac{U(z-d)}{u_* k x}\right) dx = \exp\left(-\frac{U(z-d)}{u_* k x}\right) \tag{A.1}$$

where $x[m]$ is the distance from the tower, $U[m/s]$ is the mean integrated wind speed, $z[m]$ is the measurement height, $u_*[m/s]$ is the friction velocity, $d[m]$ is the zero plain displacement height estimated as $2/3$ of canopy height h_c, and k is Von Karman constant ($k = 0.4$) [146]. $\Omega(x)$ ranges from zero at the tower base ($x = 0$) to one as x approaches infinity. In other words, for a very long distance x along the fetch and away from the tower, the cumulative contribution approaches 100%. Here, we set the maximum value of x as $x_L = 10$km, since we found that at least 90% of the flux contribution is captured within that distance (i.e. $\Omega(x_L) > 0.9$). At farther distances, the contribution from those plots is low and does not significantly change the measured flux.

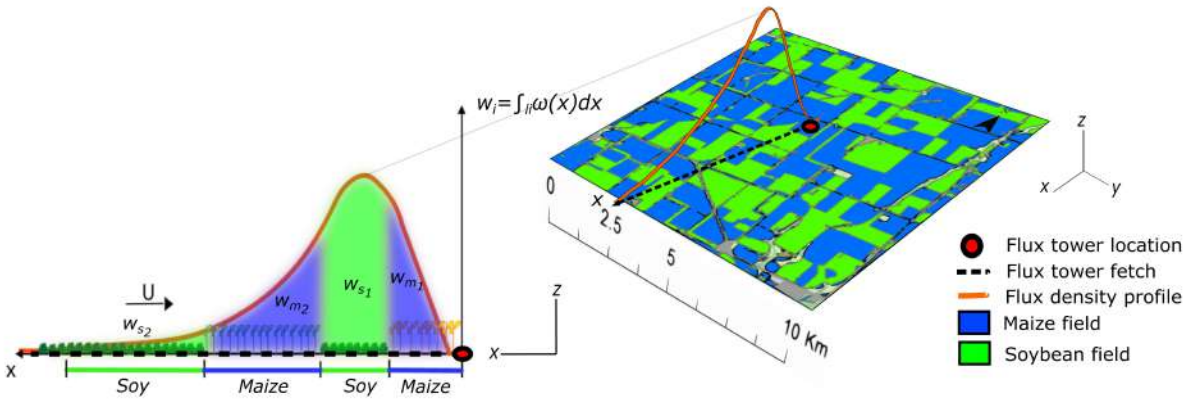

Figure A.1: Illustration of a 1D flux footprint that captures the organized heterogeneity of the maize and soybean mosaic. The land cover data from the *USDA* for the crop year 2016 (right), shows a mosaic of maize and soybean plots surrounding the flux tower. The density profile over the mosaic (orange) [6], [17] represents the relative flux contribution as a function of the upwind distance denominated as the fetch (black dashed line). The density profile shows that large contributions come from locations close to the tower, but not immediately underneath [146], and decreases as the distance in the x-direction from the tower increases. From a one-dimensional perspective, the flux tower measurements are the combined response of the fields intersected by the flux tower fetch (left). The integration of the fluxes generated by the contributing fields as a function of its distance from the tower provides the total flux.

A.1.2 Heterogeneity and flux partitioning equations using a one-dimensional flux footprint approach

Here we describe the approach to estimate the relative flux contribution due to the heterogeneous land cover when using a one-dimensional flux footprint model. Analytically, the distribution of a diffusive quantity in the lower layer of the *ABL* is described as an integral diffusion equation. Therefore, the flux footprint relates the vertical eddy flux η measured at observation height, z_{obs}, to the spatial distribution of ground source (or sink) fluxes $\mathscr{F}(x)$ at the ground ($z = 0$) at a stream-wise (x) and crosswind (y) direction from the tower location [16], [91]–[93]:

$$\eta = \int_0^{x_L} \mathscr{F}(x) \cdot \omega(x) dx \tag{A.2}$$

where $\omega(x) = \Omega'(x)$, such that it is the relative contribution to the flux at location x. $\mathscr{F}(x)$ is the source (or sink) flux from the surface at location x, with the same units as η. The observed flux η is then the weighted integration of all the surface fluxes along the fetch, and has units of W/m^2 for latent heat flux, LE, and sensible heat flux, H, and $\mu mol/m^2 s$ for CO_2 flux, F_c. In this approach, it is assumed that the turbulent flow field is horizontally homogeneous [92]. Consequently, the relative contribution due to the flux of each field source (or sink) is a function of its location away from the flux tower. In a case where the landscape consists of patches of different crops, we assume that the fluxes are the same within a given patch along the fetch, and there are n landscape patches that contribute to the measured flux. To account for a measure of surface heterogeneity, the source emission or sink rate for this number, n, of different sources at discrete stream-wise patches along the fetch can be expressed as:

$$\eta = \sum_{i=1}^{n} F_i \cdot w_i \tag{A.3}$$

where i is a patch count of length l_i along the upwind direction, F_i is the ground level flux for patch i, and w_i is the weight function of patch i as follows:

$$w_i = \int_{l_i} \omega(x) dx. \tag{A.4}$$

Therefore, the flux observed at the tower is a weighted sum of fluxes from the n landscape plots of soybean and maize located along the fetch (Equation A.3). We assumed that the modeled flux is the same for all patches of a given crop type, such that we can aggregate those patches by adding the contributions of each patch type, as follows:

$$\phi_m = \sum_{\forall i \in m} w_i \tag{A.5}$$

$$\phi_s = \sum_{\forall i \in s} w_i \tag{A.6}$$

where the subscripts m and s refer to maize and soybean, respectively. Here we assume one representative flux value for maize as C_m, and for soybean as C_s. This constitutes the assumption that the vegetation type is the dominant source of flux variability and ignores unmeasured influences such as soil moisture or nutrients along the fetch. This assumption could be relaxed if detailed characteristics of each landscape plot are available. Here we use a multi-layer canopy model to estimate the vegetation-level fluxes (LE, H, F_c) for each crop type.

Since these fluxes, here denoted as C_m and C_s, are modeled, their application to Equation (A.3) results in a total flux η_{mod}:

$$\eta_{mod} = C_s\phi_s + C_m\phi_m \tag{A.7}$$

Due to the previous assumptions, η_{mod} will not be equal to the observed flux at the tower, η, but a close estimation given the previous considerations. We compute the partitions of the fluxes measured at the tower for maize and soybean fields, respectively, as follows:

$$\eta_m = \eta \cdot \frac{\eta_{m,mod}}{\eta_{mod}} = \eta \cdot \frac{\sum(C_m \cdot \phi_m)}{\sum(C_m \cdot \phi_m + C_s \cdot \phi_s)} \tag{A.8}$$

$$\eta_s = \eta \cdot \frac{\eta_{s,mod}}{\eta_{mod}} = \eta \cdot \frac{\sum(C_s \cdot \phi_s)}{\sum(C_m \cdot \phi_m + C_s \cdot \phi_s)} \tag{A.9}$$

A.1.3 Illustration of the role of organized heterogeneity along the fetch

At a single time step, we consider that the total contributions of flux from maize and soybean along the tower fetch (ϕ_m and ϕ_s respectively) are defined by the sum of their relative contributions (Figure 2.4), as expressed in Equations (A.5)-(A.6). In an illustrative example, we consider three cases (Figure 2.4, Table A.1) where $\phi_s = \phi_m = 0.5$ (Case A), ϕ_s=0.4 and ϕ_m=0.6 (Case B), and ϕ_s=0.6 and ϕ_m=0.4 (Case C). Case A corresponds to equal contributions from the two crops as would be expected if the two were randomly (homogeneously) distributed or the organized heterogeneity incidentally reflected equal contributions. Cases B and C reflect larger contributions from maize or soybean, respectively. For all three cases, let us assume that the modeled fluxes LE_s=120 W/m^2 and LE_m=80 W/m^2, and the tower observed flux η = 105 W/m^2. For Case A this leads to a total estimated flux n_{mod} = 100 W/m^2 = (.5)120 + (.5)80. For Cases

Table A.1: Comparison of heterogeneous and homogeneous contribution of flux towards the estimation of latent heat flux, $\eta(LE)$, when using a one-dimensional approach.

Case	η(LE)	ϕ_s	ϕ_m	LE_s	LE_m	η_{mod}	η_s	η_m	$\eta - \eta_{mod}$	diff from A
	W/m^2	*frac*	*frac*	W/m^2	W/m^2	W/m^2	W/m^2	W/m^2	W/m^2	%
A	105	0.5	0.5	120	80	100	63.00	42.00	5	0
B	105	0.4	0.6	120	80	96	52.50	52.50	9	-4
C	105	0.6	0.4	120	80	104	72.69	32.31	1	4

Case A refers to the homogeneous assumption (Figure 2.4a), whereas Case B (Figure 2.4b) & Case C are heterogeneous situations that show opposing conditions in terms of dominance of soybean versus maize along the fetch, respectively.

B and C, we would similarly estimate n_{mod} as 96 W/m^2 and 104 W/m^2, respectively. Since for all three cases, the observed flux at the tower is the same, we can estimate the error as the difference between η and η_{mod} (Table A.1). In this example, we see that the landscape heterogeneity of Case C leads to the closest estimate to the flux measured at the tower. We can also consider the percent difference in η_{mod} relative to the equal contributions of Case A. Here, we see that Case B leads to a -4% difference in *LE*, due to the higher contribution from maize, which has a lower *LE*. The opposite occurs for Case C (Table A.1).

This illustrative example also demonstrates that the change in the relative contribution of fluxes due to two crop species can either increase or decrease the total flux observed at the tower (η) as a function of the field's location in the tower. We demonstrated here that an incremental change in the flux observed by a tower, η, does not necessarily correlate to an incremental change in component fluxes from a given field or crop type, but to the relative contributions of the land cover due to different arrangements.

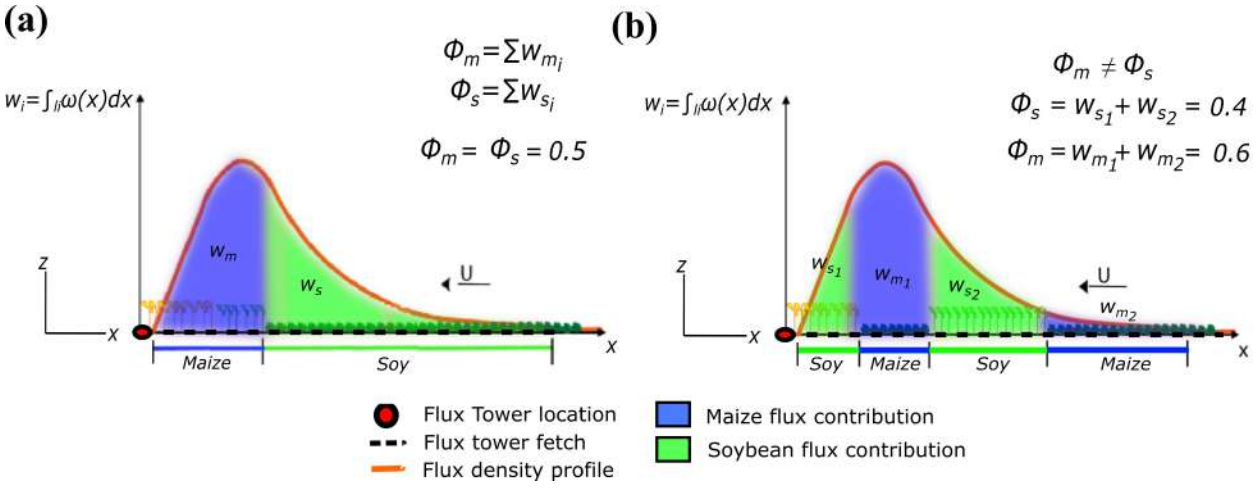

Figure A.2: Flux responses of (a) a hypothetical homogeneous mosaic, and (b) an organized heterogeneous mosaic as observed at the flux tower. In the hypothetical homogeneous mosaic, the flux attributed to maize and soybean are equal, $\phi_m = \phi_s$. Notice that equal flux response does not imply an equal distribution of fields planted with the same crop type.

A.1.4 Seasonal evolution of flux contribution due to organized mosaic heterogeneity

We estimated the daily average contribution of fluxes every 6 weeks for maize and soybean (Figure 2.10) using the observed flux tower data as described in section 2.3.3. The overall trend of LE and H exhibit a strong seasonality (Figures 2.10a and 2.10b, respectively). LE increases over the growing season, peaks around June 1-July 15, and decreases until December 1-January 15. This 6-week window analysis shows that during the 2016 and 2017 growing season, maize contributed more towards daily LE than fields planted with soybean, whereas during 2018 soybeans contributed more than maize. In terms of CO_2 flux (F_c), maize and soybean fields are CO_2 sinks during the growing seasons (Figure 2.10c). We observe a higher CO_2 release before planting (i.e., around April) (Figure 2.10c) that can be explained by rising temperatures and soil moisture that support heterotrophic respiration of existing biomass on and in the soil. Three weeks after planting, the system shifts to CO_2 sink and remains so until harvest starts. From September 1 to October 15, a transition period between positive and negative values of F_c occurs due to harvest. Right after harvest, CO_2 is not taken up by crops and a release of CO_2 is expected from crop residues due to the organic matter decomposition of stover and roots that remain in the field [98]. Finally, after harvest and before the next planting season starts, the land-surface is a net source due to soil respiration. In terms of the relative flux contribution, ϕ, (Figure 2.10d), there is a greater contribution due to soybean plots than due to maize plots.

To establish which crop was seasonally dominant in terms of a given flux, we determined the 6-week averaged difference of the percentage of contributions between the organized heterogeneous mosaic and the results from the hypothetical homogeneous case (e.g. Case A in Section 3.4) (Figure 2.11). Taking July to August 2017 window from the 6-weeks analysis as example, the contributions of flux for maize and soybean are ϕ_m=51.5% and ϕ_s=48.5%, respectively (Figure 2.10d). The vegetation-level fluxes are estimated using MLCan for soybean and maize, C_s and C_m, respectively, which combined with the relative flux ϕ_m and ϕ_s, provide the modeled contribution for each crop and the partition of the observed flux, LE_m=48W/m^2 (Equation A.8) and LE_s= 37W/m^2 (Equation A.9) (Figure 2.10a), while the averaged daily observed flux is LE_{obs}=85W/m^2. The fact that maize fields have on average a larger contribution towards LE_{obs} than soybean from July to August 2017, means that the response of the heterogeneous organized mosaic exceeded by 2% the homogeneous assumption of ϕ_m=ϕ_s=0.5 (Figure 2.11a). Similarly, on average from July to August 2017, the response of the heterogeneous mosaic resulted in 8.5% exceedance for H (Figure 2.11b) and 5% for F_c (Figure 2.11c) with respect to the homogeneous assumption. This example shows how the difference of the relative flux response due to each crop type in the organized heterogeneous mosaic can be significantly different from a homogeneous assumption, which can lead to incorrect interpretations if not considered.

In general, while maize fields contributed more to LE in summer 2016 and 2017, soybean consistently contributes higher towards total LE in 2018 (Figure 2.11a). A stronger contrast in H is generated by maize contributions (Figure 2.11b). The highest H rates were estimated during July 2017 for maize, when the lowest PPT was observed (Figure

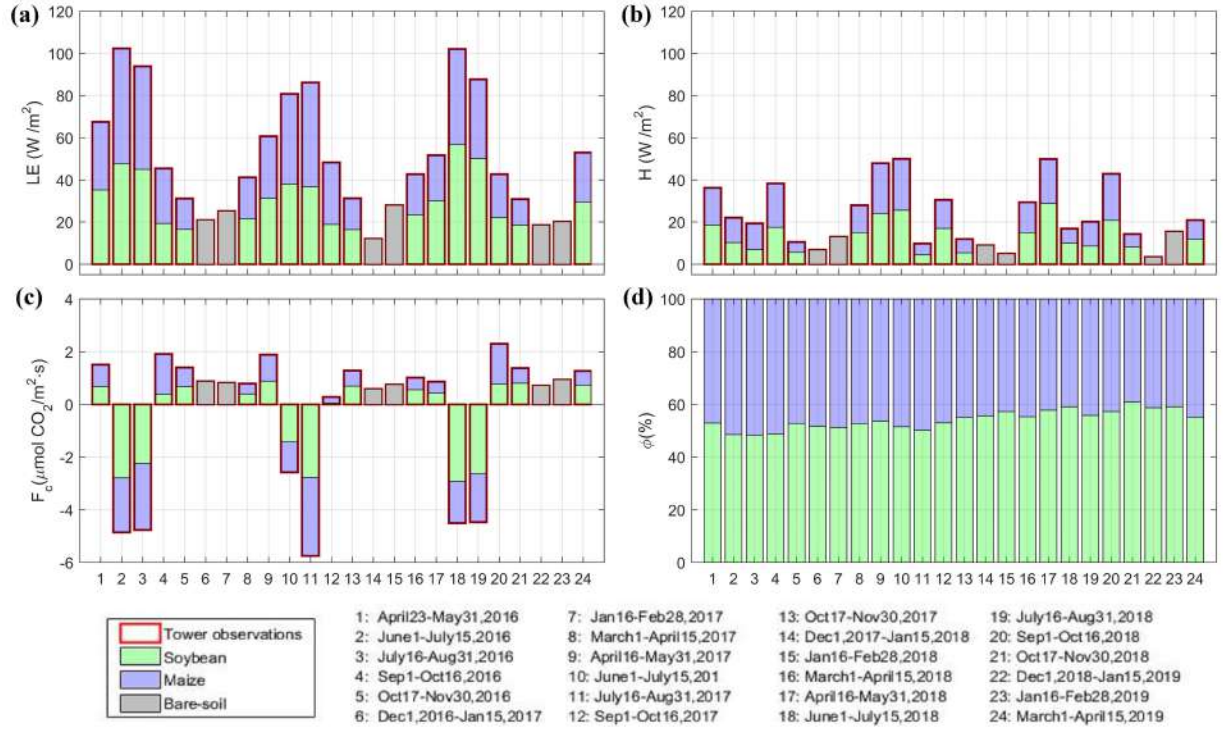

Figure A.3: Seasonal evolution of average daily fluxes due to the relative contribution of maize (blue) and soybean (green) in 6 week windows (corresponding to Figure 2.7), for (a) latent heat flux, LE; (b) sensible heat flux, H; (c) carbon dioxide flux, F_c; and (d) relative contribution of flux inside the flux footprint, ϕ. Each stacked bar refers to 6 weeks averaged daily flux for each crop type. Relative contributions due to soybean and maize were calculated with the method described in Section 2.3 for 105,216 15-minute data points from April 2016 to April 2019 obtained from the flux tower.

2.6). Overall, larger effects due to land cover heterogeneity were observed for CO_2 flux (F_c) during October-November 2018, when the soybean contribution exceeded the homogeneous assumption by 13% (Figure 2.11c), also matching the stronger contribution for latent heat flux of 6% (Figure 2.11a).

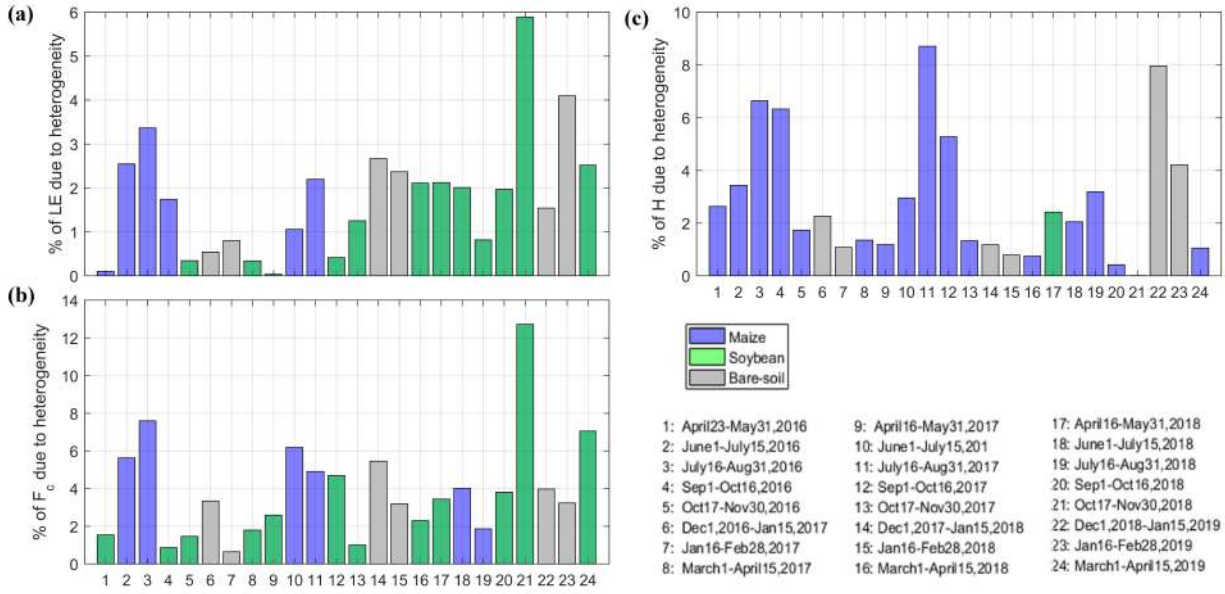

Figure A.4: Seasonal evolution of the percent contribution due to organized mosaic heterogeneity in 6-week windows (corresponding to Figure 2.7), for (a) latent heat flux, LE; (b) sensible heat flux, H; and (c) carbon dioxide flux, F_c. These plots show the difference between the heterogeneous response and a hypothetical homogeneous assumption where $\phi_m = \phi_s = 0.5$.

Some limitations and considerations apply to the previous analyses. First, given the location of the flux tower surrounded by agricultural fields, we consider that crop and soil components are the primary sources of CO_2. Low traffic in the nearby farm roads and other local sources are considered negligible contributors for this analysis. Second, the formulation employed for the footprint function (Equation A.1) is defined for near-neutral atmospheric conditions [17]. We therefore expect that fluxes are underestimated during stable conditions, such as night-time and winter, and overestimated during unstable conditions, as in summer-time when buoyancy leads to convection. Consequently, the magnitude of the observations (Figure 2.12, red outline) is different from the modeled flux based on the flux footprint partitioning. However, the percentage of the relative contribution due to each crop type enables the analysis of the relative influence of each crop type on the total flux tower observations.

A.1.5 The significance of the flux footprint method

"The footprint encompassed by a point flux measurement is the influence of the properties of the upwind source area weighted with the footprint function" [14], [92]. Here, the one-and two-dimensional flux footprint methods used in this study when estimating the role of the organized land cover heterogeneity on land-atmospheric fluxes were formulated

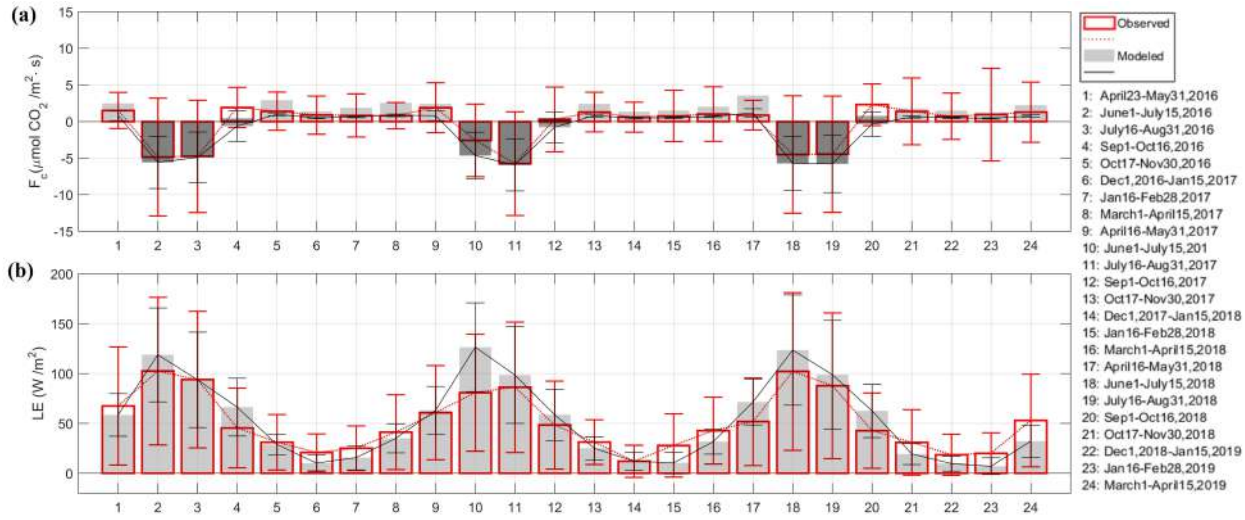

Figure A.5: Difference between flux tower observations (red) and modeled (gray) daily averaged fluxes in 6 week windows (corresponding to Figure 2.7), for (a) carbon dioxide flux, F_c, and (b) latent heat flux, LE. The error bars show the standard deviation of the daily average over the 6 week period for the observed (red) and modeled data (black). The segmented line in red shows the behavior of the daily averaged observations, while the continuous black line similarly shows the modeled data.

using different theoretical considerations. The one-and two-dimensional flux footprint, Schuepp, Leclerc, MacPherson, *et al.* [17] and Kljun, Calanca, Rotach, *et al.* [18], were selected because they allow the determination of the flux footprint from atmospheric variables obtained during flux measurements [14]. The following are the more relevant remarks of these methods:

- Dimensionality
  - The one-dimensional method proposed by Schuepp, Leclerc, MacPherson, *et al.* [17] describes the footprint function's upwind extent but not its width.
  - The two-dimensional method proposed by Kljun, Calanca, Rotach, *et al.* [18], describes both the function's upwind extent and the crosswind spread of the footprint (i.e., length and width of the flux footprint.)
- Formulation
  - The one-dimensional method proposed by Schuepp, Leclerc, MacPherson, *et al.* [17] is an analytical footprint model of the analytical solution to the diffusion equation proposed by Gash [147].
  - The two-dimensional method proposed by Kljun, Calanca, Rotach, *et al.* [18] estimate the flux footprint based on the Lagrangian stochastic particle dispersion model [148].
- Atmospheric stability

  - The one-dimensional model proposed by Schuepp, Leclerc, MacPherson, *et al.* [17] does not consider the influence of atmospheric stability when estimating the behavior of the behavior of the contribution along the flux footprint fetch. It uses a neutral stratification and the average wind velocity.
  - The two-dimensional solution proposed by Kljun, Calanca, Rotach, *et al.* [18] is valid for a wide range of boundary layer stratification's and receptor heights [18].

- Terrain conditions
  - Schuepp, Leclerc, MacPherson, *et al.* [17]'s implementation is limited to homogeneous terrain. [14].
  - Kljun, Calanca, Rotach, *et al.* [18] model can be used in heterogeneous or disturbed landscapes.

When comparing the results of the relative contribution due to each crop type when estimating the role of the organized land cover heterogeneity on land atmospheric fluxes using the one-and two-dimensional flux footprint methods, our results show that:

- A larger differentiation of the relative contribution due to each crop type is achieved using a two-dimensional approach. Moreover, we observed that using a one-dimensional method to estimate the flux footprint results in a more subtle distinction between the contribution of maize and soybeans plots (Figure A.3d) with respect to the two-dimensional approach in Chapter 2.
- We observe that the consideration of significant factors included in the two-dimensional footprint has promoted the detection of the heterogeneity of the land cover.
- Our results suggest that the relative contribution of flux reflects the scale at which the organized land cover heterogeneity is estimated. Therefore, a higher dimensional flux footprint approach has resulted in a higher estimation of land cover heterogeneity.
- We observed that the selection of a high-dimensional flux footprint model increases the estimation of the difference between the two contributing crop types when determining the role of the organized land cover heterogeneity.

Therefore, we observe how the selected representation of the flux footprint has an important role on the attribution of the relative contribution of flux due to the organized land cover heterogeneity. In this case, the more detailed version of the flux footprint produces a more detailed perception of heterogeneity and as a result, a more distinguishable relative contribution is attributed to each crop type in the 6-weeks analysis.

Table A.2: Hydroclimatic variables measured at Goose Creek flux tower site.

Instrument	Variables measured	Height from land-surface
Gas Analyzer	Ecosystem densities of H_2O and CO_2.	25 m
3-D Sonic Anemometer	U_x,U_y, and U_z wind components.	25 m
Temp and RH sensor	Air temperature and relative humidity.	25 m
Spectral Reflectance Sensor - NDVI	Normalized difference vegetation index.	10 m
Rain gauge	Precipitation.	Separate post at 1.5 m
Pyranometer	Incoming and outgoing shortwave and longwave radiation.	10 m
Infrared Radiometer	Land-surface temperature.	2 m
Soil Probe	Bulk electrical conductivity, volumetric water content, and soil temperature.	5,15,30,50,100, and 200 cm depth
Water content reflectometer	Bulk electrical conductivity, volumetric water content, and soil temperature.	Shallow depth

A.2 Hydroclimatic variables measured at Goose Creek flux tower site

At the Goose Creek eddy covariance flux tower site, more than twenty instruments are powered by an off-grid solar panel system and are located in the subsurface, near-surface, and at 10m and 25m above the surface. Table A.2 describes the instruments name, variables measured, and location at the tower.

Meteorological, soil, precipitation, radiative, and eddy covariance flux data are averaged at a temporal resolution of 15 minutes. The data is then uploaded to the *Clowder* web-based data system [149] of the Critical Interface Network for Intensively Managed Landscapes project *(CINet)* https://cinet.ncsa.illinois.edu/. The data used in chapter 2 used as forcing data for MLCan2.0 model and other data used in Chapter 2 are publicly available at GitHub (https:/github.com/HydroComplexity/MLCan2.0/tree/master/Data/Goose_Creek).

DOI: 10.5281/zenodo.3931001

The high-frequency data is measured at 25m height, it is stored locally in a data logger, and it is manually collected every month at the tower since April 2016 by Steve Sargent from the Illinois State Geological Survey. The high frequency data used in chapters 3 and 4 is standardized in a preprocessing step, by subtracting the mean and dividing by the standard deviation for each of the 3-min windows. The codes and data for the evolution of the Causal Structure used in chapters 3 and 4 are available at Github (https://github.com/HydroComplexity/CausalStructureEvolution).

For more information about the instrumentation and data at this flux tower, as well as other instrumented nearby locations, we refer the reader to Wilson, Abban, Keefer, *et al.* [27] and to the Critical Interface Network (CINet) website (https://cinet.ncsa.illinois.edu/).

A.3 MLCan parameters

The 1-D Multi-Layer Canopy model (MLCan) describes the processes of the coupled canopy, leaf, root, and soil system vertically through the canopy and soil domains [78], [94], [150]. For the simulations performed in chapter 2, we defined the soil parameters, leaf profiles, and root profiles for soybean and maize to be the same as for Bondville (Illinois) Ameriflux-site [78] given that the study site is located within a few kilometers distance. Important exceptions were made to appropriately include some particular characteristics of the Goose Creek site, such as the tower and canopy height. Some of the more relevant parameters used in our simulations are summarized in Table S1. For an extended explanation of the parameters when using MLCan model, we refer the reader to Drewry, Kumar, Long, *et al.* [78] and Le, Kumar, Drewry, *et al.* [94].

Table A.3: Parameters List for Maize and Soybean Used in MLCan

Description	Symbol	Units	Maize	Soybean
Photosynthesis			C4	C3
Maximum rate of Rubisco carboxylation	$V_{cmax_{25}}$	$\mu mol\ /m^2\ s$		100
Substrate saturated Rubisco capacity	V_{max_4}	$\mu mol\ /m^2\ s$	40	
Maximum electron transport rate	$J_{max_{25}}$	$\mu mol\ /m^2\ s$	180	
Leaf respiration	R_d	$\mu mol\ /m^2\ s$	0.8	1.5
Conductance and leaf physiology				
Ball-Berry slope	m	-	7	10.6
Ball-Berry intercept	b	$mol\ /m^2\ s$	0.008	0.007
Leaf potential at which half of the hydraulic conductance is lost		m	-2	-2
Leaf and canopy				
Leaf width (needle diameter)	d_o	m	0.08	0.06
Canopy height	h	m	2.5	1
Flux tower observation height	h_{obs}	m	25	25
Soil property				
Percent of sand		%	5	5
Percent of clay		%	25	25
Root structure				
Maximum root depth	r_d	m	1.5	1.5
50^{th} percentile rooting depth	z_{50}	m	0.28	0.28
95^{th} percentile rooting depth	z_{90}	m	0.92	0.92
Litter Layer				
Litter layer depth		m	0.03	0.03

A.4 LE & F_c coupled behavior

Given the dependence between latent heat (LE), and carbon dioxide flux (F_c), we explored their coupled behavior during the growing season. Maize and soybean were simulated using MLCan model [78], [94], [150]. We observed that

the four stages -planting, growing, maturing, and harvest- cluster together for maize and soybean (Figure A.6). During planting and harvest, lower LE is observed, whereas LE peaks during the mature stage. CO_2 uptake (negative F_c) peaks for maize and soybean during the growing stage (Figure A.6, orange circle). Maize is harvested later than soybean, which corresponds to continued negative F_c for maize from September to mid October and positive F_c for soybean.

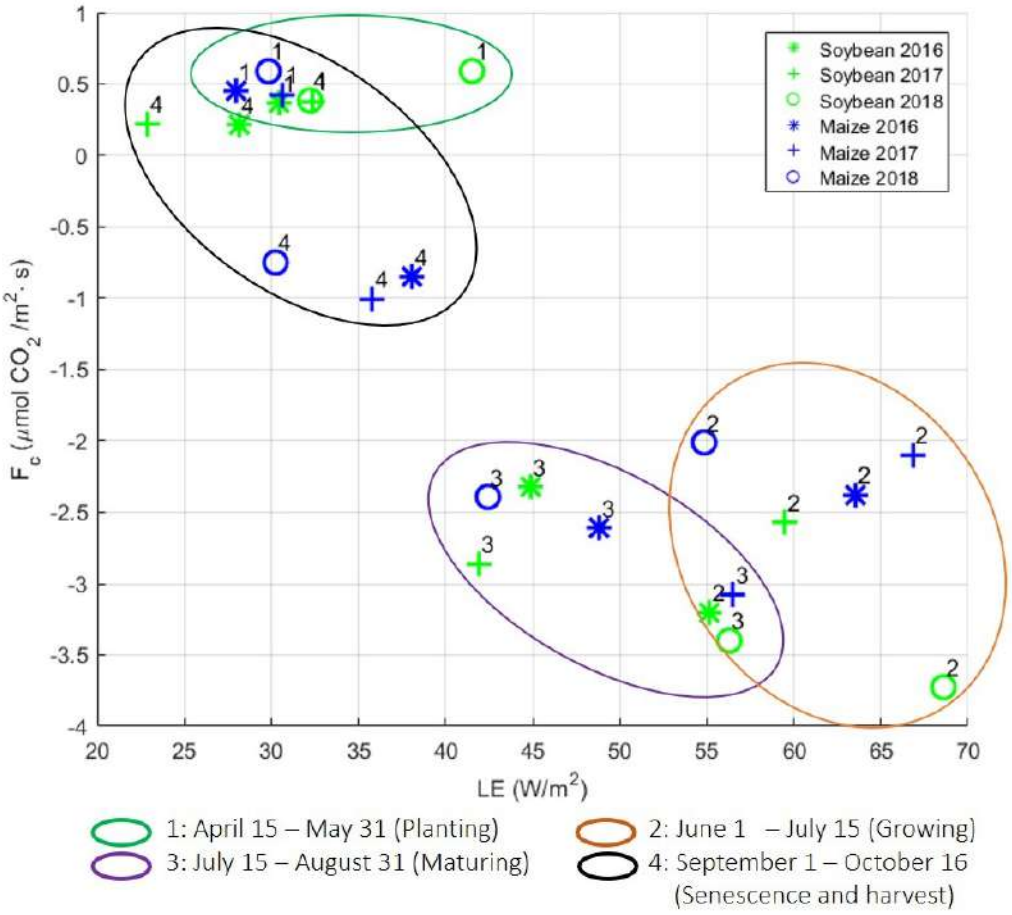

Figure A.6: Modeled latent heat flux (LE) versus carbon dioxide flux (F_c) for soybean and maize through four stages of the growing seasons 2016 - 2018. The numbers in the figure indicate crop's stages: (1) Planting, (2) Growing, (3) Maturing, and (4) Senescence and harvest. The colored circles show the clusters generated by the stages.

A.5 A toy example of using the $K\,(G_l, G_m)$ metric

Here we provide an example for the estimation of the distance metric, $K\,(G_l, G_m)$, used to compare two DAGs proposed by Malmi, Tatti, and Gionis [115], which we have implemented for DAGs of ecohydrological time series data. For a pair of toy DAGs, $G_l = (V, E1)$ and $G_m = (V, E2)$ (Figure A.7). Given that the two DAGs are not identical, meaning that they share the same nodes but not the same edges, using Equation 3.3 we can expect to obtain a distance K

larger than zero.

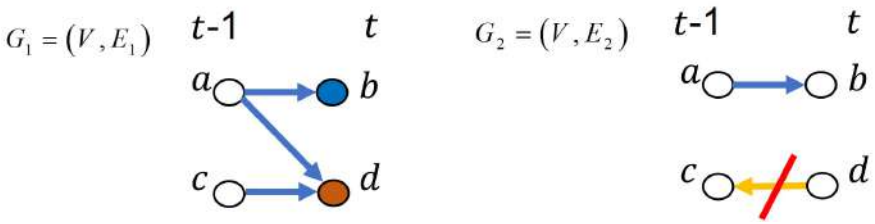

Figure A.7: A pair of toy DAGs related to the example in Appendix A.5, based on Example 1 in Malmi, Tatti, and Gionis [115] when implemented for DAGs for time series.

To compute the distance between the toy DAGs shown in Figure A.7, we start counting for the number of concordant edges between them. One concordant edge, namely, (a, b), connects the same nodes in both DAGs, then K=0. Next, there is one discordant pair, (c, d) in G_l and (d, c) in G_m as defined by Malmi, Tatti, and Gionis [115], which cannot exist in the DAGs for time series of instantaneous data given that only the past can have a causal influence on the future. One pair that is not in the other graph, namely, (a, d) results in K=1, only if p in Equation 3.3 is $p = 1$. Now, we look for feasible edges that are missed in both graphs. Given that only non-contemporaneous edges that represent lagged in time influence can be missed, there is only one edge missed in both graphs which is (c,b). The rest of the missed edges are contemporaneous edges and we do not account for them in DAGs from time series. Therefore, if $q = 1$ in Equation 3.3, then K=1. The final distance between the pair of toy DAGs is the sum of all previous distances, resulting in K=2. Notice that to avoid the risk of accounting for the nonexistence of contemporaneous links in both DAGs, a more conservative view is to set $q = 0$. In that case, the difference between the two DAGs is K=1.

A.6 Testing the performance of the $K\,(G_l, G_m)$ metric

To test the performance of the $K\,(G_l, G_m)$ metric, we evaluated two scenarios. First, we estimated the distance between identical toy-DAGs, $K\,(G, G)$, for different values of p and q (Figure A.8a). In this case, p can take any value without changing the distance, since the same link is removed from both DAGs. On the other hand, q adds to the distance for each link that is removed in both DAGs. In the second scenario, we calculated the distance between a fully connected DAG, G_{full}, and a DAG while its edges have been removed increasingly, for different values of p and q (Figure A.8b). p accounts for the difference in edges between the fully connected DAG and the partially connected DAG, increasing the distance as a linear function of the number of different edges between the DAGs. On the other hand, q is not relevant since the same edge is not being removed from the DAGs simultaneously.

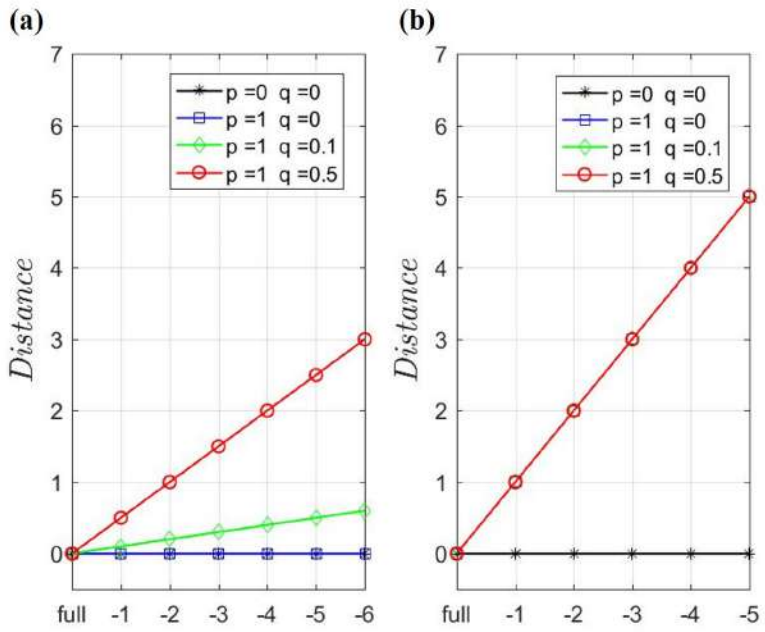

Figure A.8: Testing the distance measure for pairs of (**a**) identical toy-DAGs, $K\,(G, G)$, while simultaneously removing the same edges in both DAGs, and (**b**) a fully connected DAG and a DAG which edges are removed increasingly, $K\left(H_{full}, G\right)$. We tested 4 different combinations of p and q values.

A.7 Estimating G_{full} without contemporaneous links

An important characteristic of the DAGs from time series is that they do not have any contemporaneous links. The reason is that a variable at the current time can only be influenced by the past, whether by itself at a lagged time, by the previous state of another variable, or by its whole causal history. Therefore, when estimating a DAG from a set of turbulent time series, an edge runs from the vertex i to j, only if $i < j$ in time. Then, the maximum number of edges in a fully connected DAG with no contemporaneous links (i.e., $e_{G_{\text{full}}}$ in Equation 3.13), is determined by:

$$e_{G_{\text{full}}} = n\frac{m(m-1)}{2} + n(n-1)\frac{m(m-1)}{2} \tag{A.10}$$

where n refers to the number of variables in the DAG, while the number of time-step evaluated, including the current time, are denoted by m. An example of a fully connected DAG for n=5 and m=11, is shown in Figure A.9.

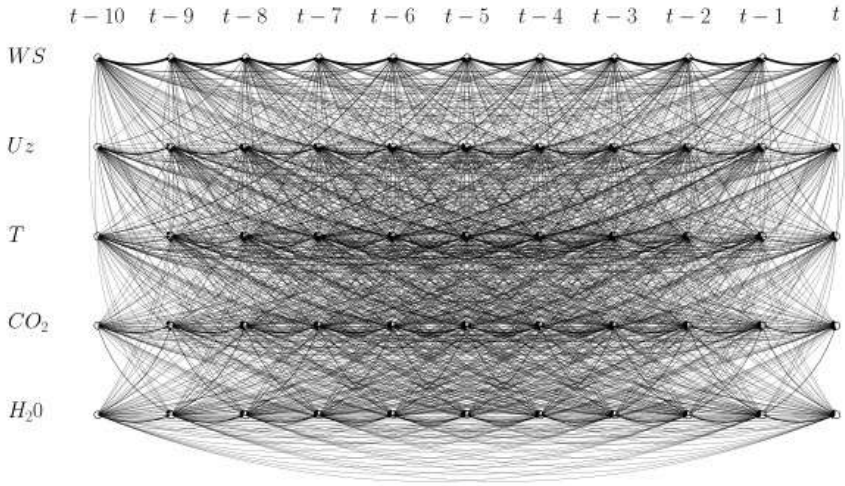

Figure A.9: Fully connected DAG from turbulent time-series, for five variables (y-axis) and eleven time-steps.

A.8 Cross-covariance of the time series in the DAGs of the cluster's centroids

Here we determine the cross-covariance of the discrete time series corresponding to the centroids of the clusters shown in Figure 3.7. In addition, we include the cross-covariance plots for the time series of 2017/08/21 12:19-12:22 CST (Central Standard Time), to determine the behavior of the self-dependence of CO_2, which is not observed in the causal analysis. The cross-covariance measures the similarity between a time series and a shifted (lagged) time series as a function of the lag. For any two variables X and Y, if the cross-covariance is positive then larger values of X are associated with larger values of Y and smaller values of X are associated with smaller values of Y. If the covariance is negative, the opposite holds: small values of X are associated with larger values of Y and vice versa. We use the cross-covariance here to determine the time lags for which a pair of time series have any type of association, causal or not, as a function of time lag, τ. We use the generated plots to complement our causal analysis when using DAGs for causal inference in Subsection 3.3.2.

We estimate the cross-covariance up to a time lag $\tau = 200$, which corresponds to 20-seconds for time series of 10Hz (Figures A.10-A.15, blue line). we normalize the sequences so that the autocovariances at zero lag equals 1. To determine the confidence intervals for the cross-covariance (Figures A.10-A.15, red lines above and below the zero line in black), we estimate the lower and upper 95% confidence bounds for the normal distribution $N(0, 1/L)$, whose standard deviation is $1/\sqrt{L}$. For a 95%-confidence interval, the critical value is $\sqrt{2}\,\mathrm{erf}^{-1}(0.95) \approx 1.96$ and the confidence interval is:

$$\Delta = 0 \pm \frac{1.96}{\sqrt{L}} \tag{A.11}$$

where L is the length of each time series, in our case 1800. In the following plots, the cross-covariance values outside of the 95%-confidence interval (Figures A.10-A.15, red lines above and below the zero line in black) refer to time series that share a significant cross-correlation at different time lags. On the opposite, the time series are not significantly correlated for the time lags where the cross-covariance is inside of the 95%-confidence intervals.

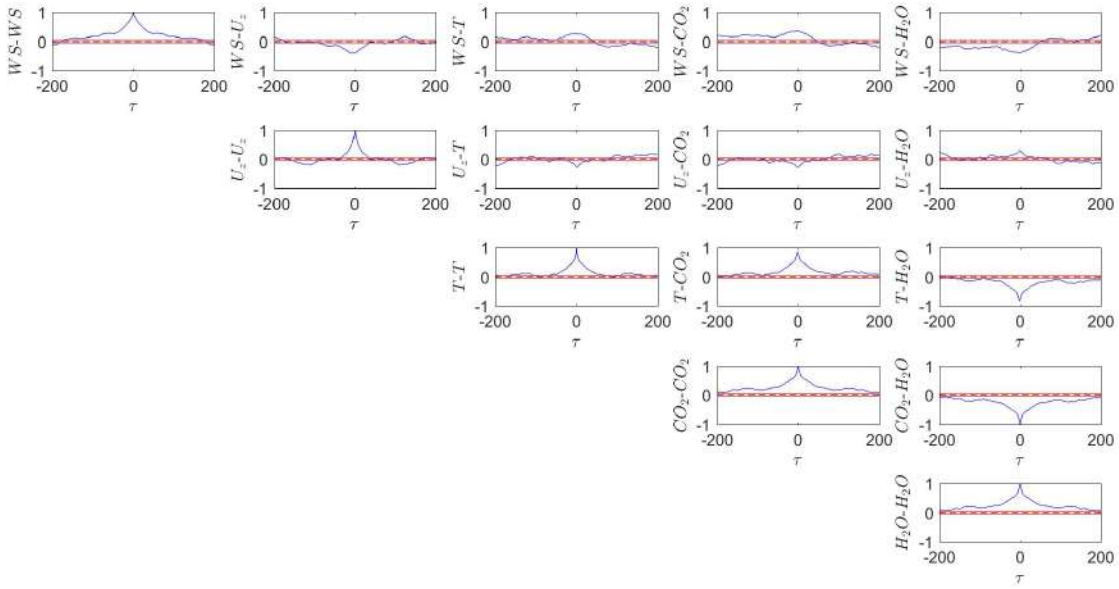

Figure A.10: Cross-covariance for 2017/08/21 13:40-13:43 CST, Cluster #3.

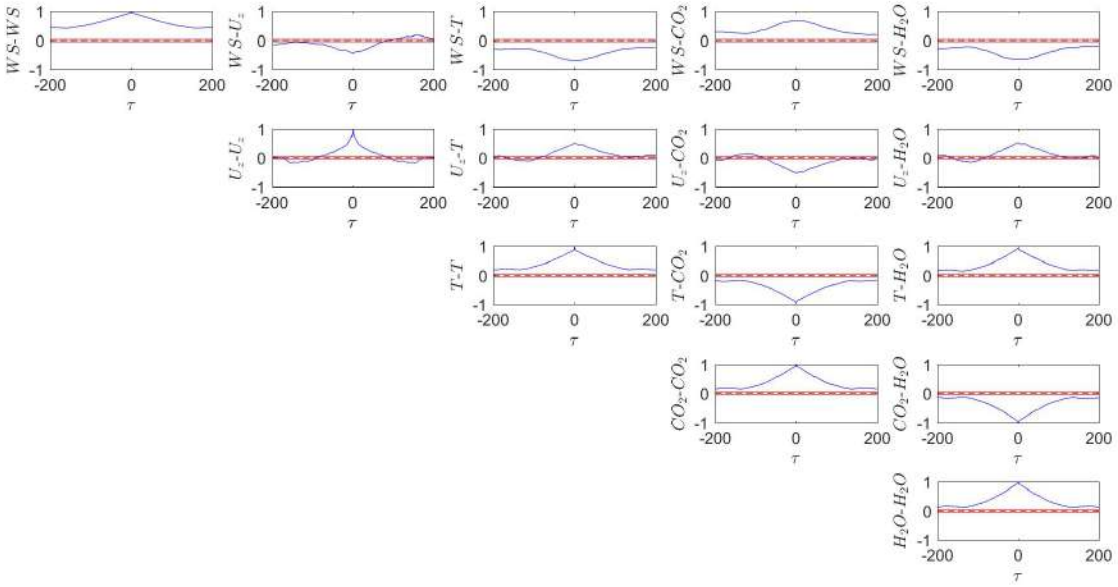

Figure A.11: Cross-covariance for 2017/08/18 10:46-10:49 CST, Cluster #3.

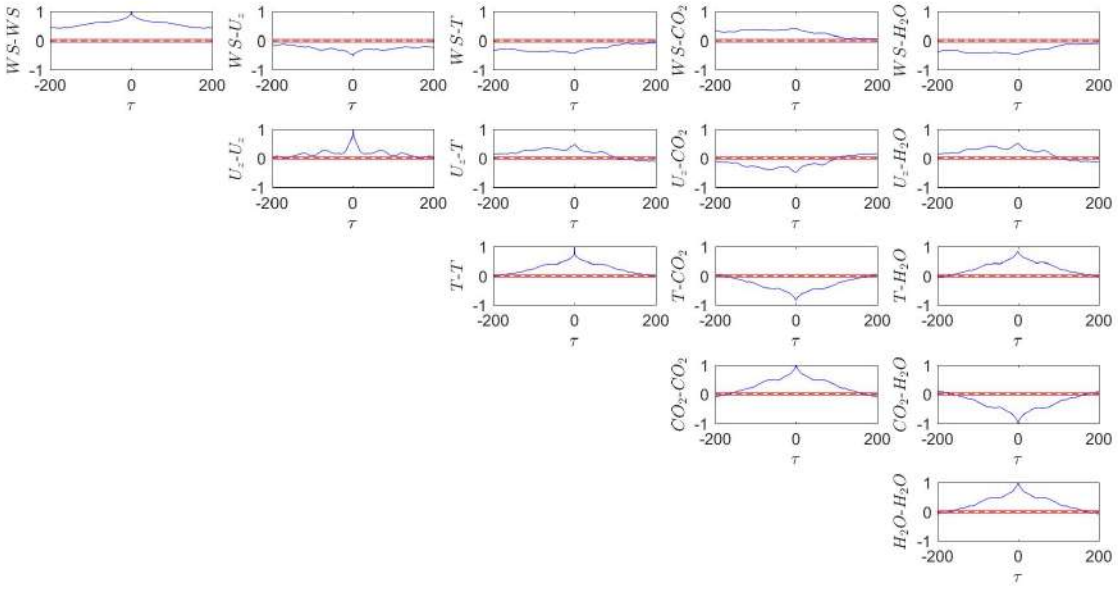

Figure A.12: Cross-covariance for 2017/08/18 10:49-10:52 CST, Cluster #2.

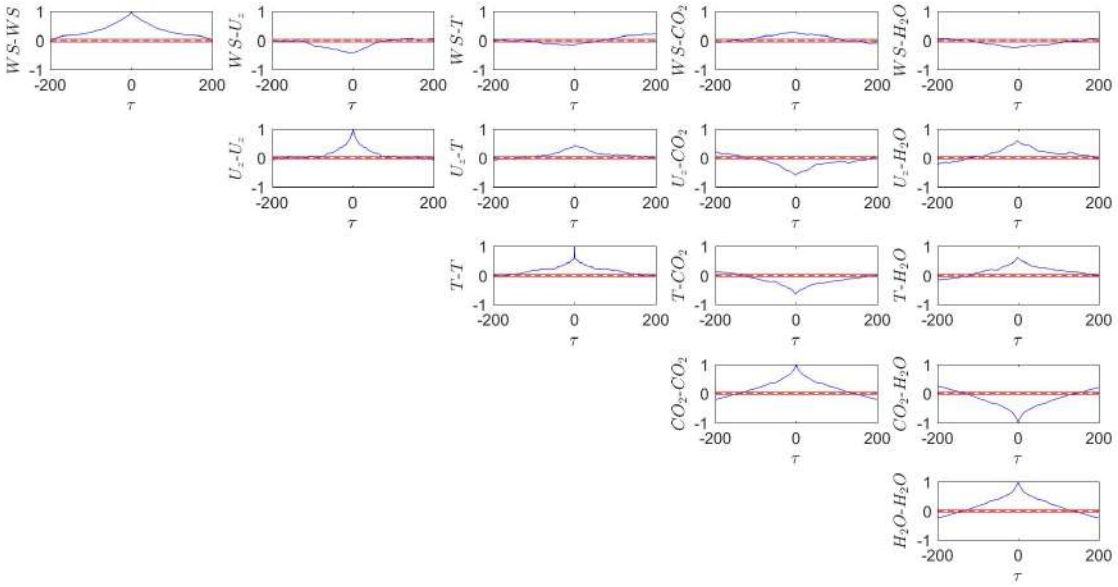

Figure A.13: Cross-covariance for 2017/08/21 12:58-13:01 CST, Cluster #2.

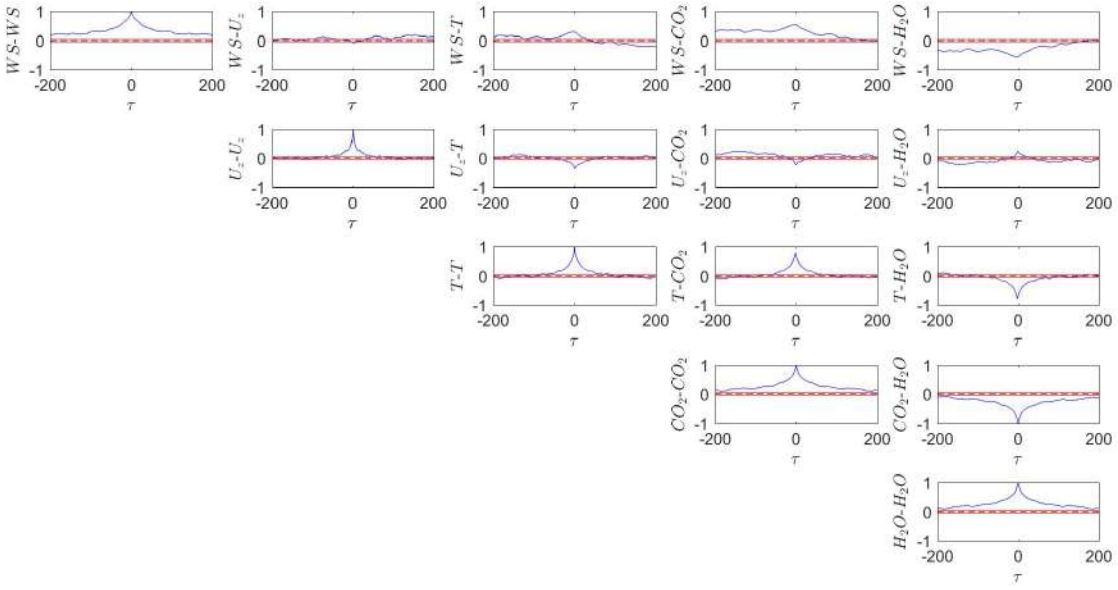

Figure A.14: Cross-covariance for 2017/08/21 13:25-13:28 CST, Cluster #1.

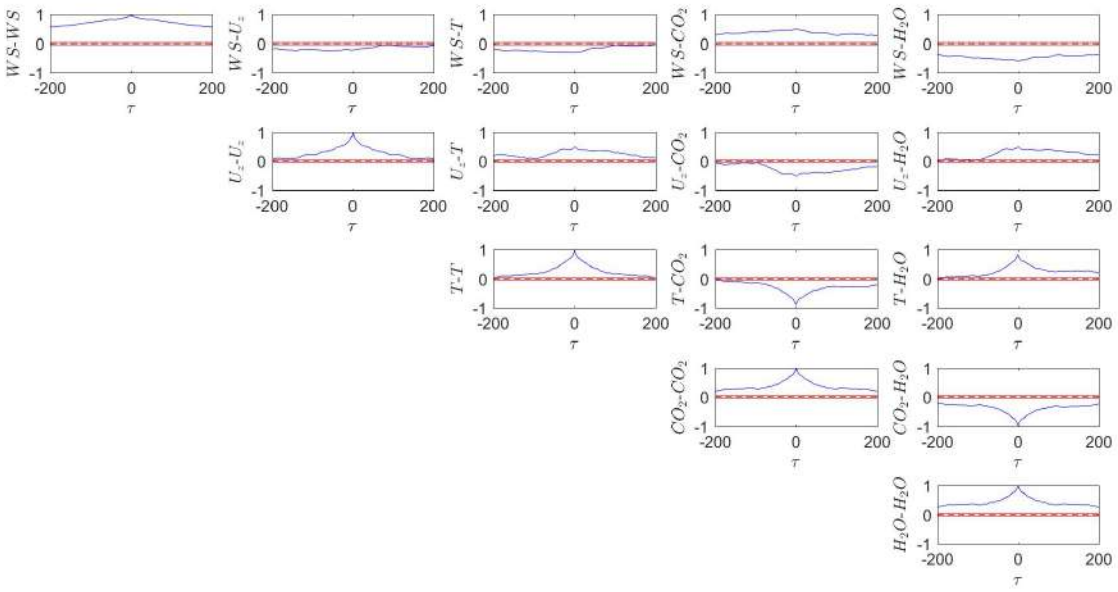

Figure A.15: Cross-covariance for 2017/08/18 12:01-12:04 CST, Cluster #1.

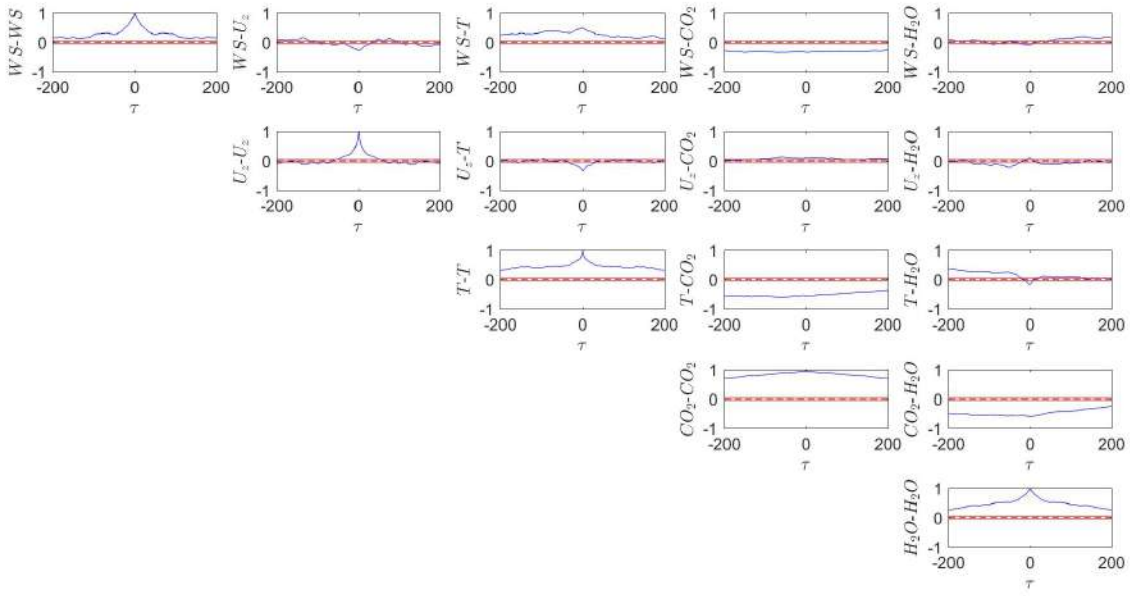

Figure A.16: Cross-covariance for 2017/08/21 12:19-12:22 CST. CO_2 shows a strong instantaneous auto-covariance at time zero that slowly die out as the time lag increases. This is a unique behavior for CO_2 that is not observed for other windows of analysis.