Information Dynamics for Complex Ecosystem Prediction and Design

A Dissertation
Submitted to the Graduate School of
Information Science and Technology
Hokkaido University

Jie Li

May 2021

Abstract

An ecosystem is a complex assembly of an uncountable number of living organisms, physical components of the environment and all interrelationships in a particular unit of space. Healthy ecosystems are "balanced" systems in which interactions among components contribute to a certain stable state of ecosystems, ensuring steady requisite ecological services for living organisms. Nevertheless, ecosystems are always exposed to variable disturbances such as the fluctuations of environmental factors, alien species invasion and internal diseases and disorders, that may influence the structure and function of ecosystems and result in ecosystem degradation and biodiversity loss. Ecosystems are highly dynamical and nonlinear complex systems, making it challenging to monitor, understand and regulate adequately. In recent decades, data-driven network approaches and mathematical analyses have been increasingly used in ecosystems research thanks to their visualization, simplicity and analyzability. Complex ecosystems are therefore abstracted as a set of nodes representing individual species and environmental factors and a set of links characterizing biotic and abiotic interactions among these living and nonliving components, forming the formalism of graph with particular structure and function. Therefore, methodological frameworks in graph theory can be well exploited to investigate the dynamics and stability of ecosystems, and recognize species-specific features and collective behavior.

In this research, complex network models are used to disentangle the complexity of ecosystems, and study the information dynamics and dissemination among components. Information-theoretic variables including transfer entropy, mutual information and Shannon entropy are incorporated into complex networks, formulating an integrated Optimal Information Flow (OIF) model. When inferring complex networks for ecosystems, the detection of interrelationships between components is one of the fundamental work for ecosystem modeling and graphical representation. The proposed information-theoretic OIF model quantifies these interrelationships

by measuring causal interactions that can be perceived as information fluxes. The performance of OIF in inferring causal interactions is validated on a mathematically simulated predator-prey model, a real-world sardine-anchovy-temperature system and a multispecies fish community by comparing to the well-documented Convergent Cross Mapping (CCM) model. Results from the validation work demonstrate that OIF outperforms CCM since it provides a larger gradient defining causal interactions at higher resolution, smaller fluctuations, more accurate prediction for ecological indicators and no requirement for convergence. Thus, the proposed OIF can be used as a robust model to infer causal interactions and networks in ecosystems. The information-theoretic causal interactions should be considered here as nonlinear predictability of ecological information about species communities.

This research also explores OIF's applications in two real-world ecosystems: gut microbes and a marine fish community. The gut-associated microbiome is an extremely complex ecosystem considering the large number of bacteria and their interactions. In this case study, to untangle the complexity of human microbiome for the Irritable Bowel Syndrome (IBS), OIF is used to infer species interaction networks for healthy, transitory and unhealthy groups. It is observed that healthy networks are characterized by a neutral patterns of species interactions and scale-free topology versus random unhealthy networks. The top ten interacting species are the least relatively abundant for the healthy micriobiome and the most detrimental. These results are useful for public health and disease diagnosis and etiognosis, as well as the personalized design treatments and microbiome engineering. In the case study of the marine fish community, to study the biological responses of the ecosystem to global ocean warming caused by climate change, OIF with Kernel estimator is employed to infer species interaction networks for the fish ecosystem considering five temperature ranges: $\leq$10°C, 10-15°C, 15-20°C, 20-25°C, $\geq$25°C. OIF-inferred networks present different patterns in structure and function for each temperature range that indicate the evolution of system dynamics with the change of sea temperature. Network-based species-specific analysis is also performed to identify critical species that have more impacts on the fish community, and species more sensitive to the fluctuations of sea temperature. This work provides a data-driven tool for analyzing and monitoring fish ecosystems under the pressure of ocean warming and is valuable to formulate accurate fishery policy to maintain fish ecosystems stable and sustainable.

Acknowledgments

I would like to thank my esteemed supervisor Dr. Matteo Convertino for his invaluable supervision and support during my PhD program. My gratitude extends to the GI-CoRE Global Station for Big Data and Cybersecurity for the funding opportunity to undertake my studies at the Graduate School of Information Science and Technology, Hokkaido University. Additionally, I am deeply grateful to Prof. Ohgane, Prof. Miyanaga and Assoc. Prof. Tsutsui for their help in the past three years. I also thank my team, Dr. Louis Chan and Elroy Galbraith for valuable discussions on my research, and my friends, Dr. Ruihe Yang and Zhiqing Chen for the cherished time we spent together in Hokkaido. My appreciation also goes out to my family for their encouragement and support all through my studies.

Publications

Articles

1. Li J., Convertino M., "Optimal Microbiome Networks: Macroecology and Criticality," *Entropy*, 2019; 21(5): 506.

2. Li J., Convertino M., "Inferring Ecosystem Networks as Information Flows", *Sci Rep* 11, 7094 (2021).

3. Li J., Convertino M., "Temperature-driven Organization of Fish Ecosystems and Fishery Implications", 2021. (Submitted to *PLOS ONE*, under review).

4. Galbraith, E. Li, J., and Convertino M., "In.To. COVID-19 Socio-epidemiological Co-causality", 2021. (Submitted to *Royal Society Open Science*, under review).

Conference Paper without Referee

1. Jie Li, Matteo Convertino, "Inference of Complex Microbiome Networks: Macroecology and Entropy Balance," Proceedings of 2018 Summer International Symposium on Big-Data, Cybersecurity and IoT, August 7-8, 2018.

2. Jie Li, Matteo Convertino, "Taming Network Inference: Optimal Transfer Entropy Model," Proceedings of 2018 Winter International Symposium on Big-Data, Cybersecurity and IoT, December 20-21, 2018.

3. Jie Li, Matteo Convertino, "Model vs Data Centrality: Probing Transfer Entropy," Proceeding of 2019 Summer International Symposium on Big-Data, Cybersecurity and IoT, August 8-9, 2019.

Patent

1. Elroy Galbraith, Matteo Convertino, Jie Li, Victor Del Rio-Vilas. “InTo (Infodemic Tomography) COVID-19: Social-epidemiological Co-causality”, US Patent, Application Number: 63119650.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

An ecosystem is a prototypical complex system that is comprised of millions of living organisms and physical components of the environment (temperature, precipitation, the climate, for instance), interacting with each other. Ecosystem complexity can be defined on multiple dimensions: spatial heterogeneity, organizational connectivity and historical contingency [1], and develops from the vast number of species or communities and their interrelationships [2, 3]. Understanding the dynamical complexity and non-linear dynamics of ecosystems is the matter of increasing importance [4], while it has turned out to be a challenging task. Conventional ecological knowledge and approaches to study and understand ecosystems are to use data sets from experimental observations and census to track species-specific changes individually and identify possible causes for the changes considering environmental data. Although this type of analysis is able to monitor the fluctuations of species abundance, biodiversity and populations over time, it is hard to further understand internal mechanisms that drive the dynamical evolution of ecosystems [5]. In fact, biological or ecological analyses only on temporal scale are far from adequate to describe ecosystems completely. Therefore, to study the stability and sustainability of ecosystems at a system level, and to identify critical species that are most responsible for the system stability and sustainability, and dominant environmental factors having the greatest impacts on the ecosystem, it is essential to carry out both system-level and species-specific analyses considering multiple scales, and to apply interdisciplinary approaches including but not limited to mathematical models and modern information technologies to ecosystem issues.

Studies on ecological networks and community ecology emphasize the importance to study individual species and their interactions simultaneously [6–9]. For

this purpose, complex network models are a set of potent tools and algorithms that intuitively fit for ecosystem modeling. Basically, network is composed of a set of nodes, describing species or coarser functional communities, and a set of edges (links) characterizing interactions between components in ecosystems [10]. These interactions can be biotic (interactions among species themselves) and abiotic (interactions between species and external environment) [11]. In fact, ecosystem complexity is an interdisciplinary research field that borrows tools and concepts from complex network science (self-organization, criticality, and phase transition, for instance) [12]. Complex network models offer new perspectives for ecosystem research [10, 13]. First, network modeling provides a mathematical way of simplifying complex ecosystems, making it possible to visualize and observe major components and their connections. Second, network-based measures and algorithms in graph theory [14, 15] can be well used to analyze ecosystems. Third, complex network models treat ecosystems as organized systems. The principle of system stability and dynamics can be incorporated into ecosystem investigation. Fourth, network approaches allow to analyze the asymptotic collective behavior of species or coarser communities and predict the taxonomic diversity and stability of ecosystems. Additionally, the exploration of large-scale networks consisting of overwhelming numbers of nodes and interactions is now feasible thanks to the advances of graph theory and the availability of massive amount of computational power, even though the scale of network-based computation is exponentially proportional to the number of nodes.

The detection of species causal interactions (causal inference) is of critical importance for network reconstruction and representation. The first conceptualization of causality that can be computed was introduced by Wiener in 1956 [16], while for quite a long time, correlation measuring linear similarity between two variables had been considered as a quantification for causal relationships even though George Berkeley and Pearson suggested that correlation did not necessarily and sufficiently imply causation [17, 18]. Especially for ubiquitous nonlinear processes, applying linear correlation to infer causal interactions is cursory and risky. Intuitively, there are three main aspects of thinking about the phenomenon of which one variable causes another: (1) the source variable (cause) is able to predict the target variable (effect); (2) the target variable to a certain extent helps to estimate the states of source variable with time lags, given that the source variable is encoded in the target variable; (3) the source variable passes certain types of information to the target variable, leading to three main computational approaches frequently used to infer

causal interactions: Granger Causality (GC), Convergent Cross Mapping (CCM) and Transfer Entropy (TE).

Granger Causality was formalized by Granger based on Wiener's idea in terms of autoregression and predictability [16, 19]. According to the concept of GC model, a variable is said to "GC cause" another variable if the historical states of the first variable helps in predicting those of the second variable. This notion of causality was substantially based on the predictability of time series, although strictly speaking, Granger causality is about conditional independence of variables rather than predictability. The key requirement of GC model is separability that is a feature of purely linear and stochastic systems [20], and provides a way to understand the system as the sum of its parts rather than as a whole non-linear entity difficult to separate. Separability means that the second variable can be independently and uniquely forecasted by the first variable. This is an assumption that reflects how systems are interpreted as linear systems, and that is certainly not the case of real complex systems. Additionally, states in the past of some variables in dynamical systems can be inherited through time, which means that the behavior of dynamical systems has memory. Yet, both cause and effect are embedded in a non-separable higher dimension trajectory. Space-time separability therefore becomes extremely hard to satisfy in systems, which can be described as complex networks where each node (variable) influences several nodes or even all nodes in the entire system simultaneously, resulting in a non-random propagation of information through the network. In this sense, ecosystems can be thought as information machines where separability is only possible by fixing thresholds of significance for the patterns to investigate. As a consequence, GC model might be problematic while using in nonlinear dynamical systems with deterministic settings and weak to moderate interactions.

To solve causal inference problem in complex ecosystems, Sugihara et al. (2012) developed the Convergent Cross Mapping (CCM) based on empirical dynamic modeling [20], and applied this model to a coupled non-linear mathematical predator-prey model and a real-world sardine-anchovy-temperature ecosystem. Later on, Ushio et al. (2018) [21] applied CCM to a fish ecosystem involving 15 species. The seasonality of abundance data was removed to assess "true" or biological interactions. In a dynamical system, two variables (X and Y for instance) are causally linked if they are generated by one system and share a common attractor manifold. It implies that each variable can be used to recover (predict) the other one. CCM is the method capable of quantifying this kind of relationship between two variables.

CCM does so by measuring the extent to which the states of one variable (considering values rather than probability distributions) can be reliably estimated by the other one with time lags. In practice, CCM takes the time series of variables X and Y and their lagged coordinate embedding to measure the ability to estimate the states of variable X from the time lag embedding that quantifies how much signature of X is encoded in the time series of Y. This principle was termed as "Cross Mapping", and it was suggested that the causal effect of X on Y is determined by how well Y "cross maps" X. Sugihara et al. (2012) [20] noted that CCM had drawbacks, although some of these are disputable. For instance, the phenomenon of "generalized synchrony" as a result of exceptionally strong unidirectional causation (X strongly "cross maps" Y, but Y does not causes X), both directions (X "cross maps" Y, Y "cross maps" X) of the causal relationship can be observed from CCM's results, resulting in a "misleading" bidirectional causality [22]. This was perceived as a limitation of CCM in distinguishing between bidirectional causality and strong unidirectional causality because of the synchrony. Misunderstanding is however not a correct definition since we believe any variable has always non-zero interdependencies due to unaccounted factors and chance that interactions may appear at least once in the ecosystem considered. Yet, asymmetrical interdependence is a norm rather than a numerical artifact. Another key property, and potentially a drawback, of CCM is convergence which means that the estimation skill can improve with the increase of the length of time series. However, datasets are not always long enough especially for real-world applications. Yet, convergence might be limited by the finite size of time series data. Lastly, CCM suffers from the high computational complexity in terms of model parameters and computation speed. Despite these drawbacks, CCM is used in this study as a benchmark for evaluating our proposed model.

Information-theoretic variables have been increasingly used in complex networks research. TE coined by Thomas Schreiber [23] is an information-theoretic variable measuring the asymmetrical bidirectional information transfer (vs. information flow as in Lizier et al. (2010) [24] when conditional entropies are used to exclude indirect pairs of species whose interactions are of second order importance) between two variables [25]. In a conceptual and practical view, besides GC and CCM, TE can be an appropriate candidate to infer the causality between interacting elements in complex ecosystems. Given that GC model may be problematic in complex systems due to highly nonlinear dynamics, and CCM may not be suitable for distinguishing well bidirectional or strong unidirectional interactions and re-

quire convergence and high computational power, TE, as a non-parametric, model-free information-theoretic variable defined from the nonlinear dynamics of Markov chain processes (mappable as stochastic pdf propagation equivalently), provides a directed measure to detect asymmetrical dynamical information transfer between two time-varying variables. TE is particularly convenient because of no assumption of any particular functional process or numerical model to identify interactions in complex systems [26]. TE has been widely used for causal inference, general principles, unified frameworks, and models, while systematic developments for causal inference based on TE are still lacking. More importantly, no work has been done to give validations for TE-based causal inference models with mathematically synthetic data, as well as real-world ecosystems, to elucidate how TE behaves dependent on dynamics and complexity. Razak et al. (2014) [27] made progress on this issue by using classical and amended Ising models which are mathematical models of ferromagnetism in statistical mechanics; Duan et al., (2013) [28] provided a theoretical and experimental systemic validation of a TE-based model; and finally Runge et al. (2018) [29] explored TE and other models with synthetic data. However, these studies were applied to complex systems with a limited number of variables or whose dynamics is well defined; yet, they did not validate the model for realistic ecosystems in its full complexity, driven also by data fallacies, as seen in the nature. Therefore, the rigorous performance assessment of TE-based models in specific applications remains elusive. On one side, well-known ecosystems with low complexity can validate the inferred pairwise interactions, while highly complex ecosystems can validate the predictability of whole systemic interaction network on some patterns' metrics such as biodiversity. The former problem deals more with accurate causality between pairs, while the latter deals more with ecosystem predictability.

In this study, to formalize the "causality" from the perspective of predictability and information dynamics, an Optimal Information Flow (OIF) model based on TE is developed by our group [30, 31]. In particular, OIF was improved with respect to Li and Convertino (2019) [31] by considering its extension over time to reconstruct dynamical information networks, the varied Markov order of each variable and a more refined pattern-oriented criteria to select optimal threshold based on the maximization of Mutual Information. OIF overcomes the limited TE for the uncertainty reduction scheme (see Li and Convertino (2019) [31]), for the consideration of maximum information/entropy via considering the full network entropy pattern (as in Servadio and Convertino (2018) [30]), and the MI-based maximization cri-

teria to define the interaction threshold to accurately predict system patterns (e.g. biodiversity). It should be noted that the optimal threshold on interactions is not necessarily within the scale-free or maximum entropy range of inferred collective behavior. In this way, processes (interactions) are clearly linked to ecological patterns (e.g. α diversity) which provides relevance to the inference problem. OIF is dependent on the (automated) choice of appropriate time delays among variables. Furthermore, the performance of the proposed OIF model is tested on a data set mathematically generated by a predator-prey model, as well as real-world data sets of sardine-anchovy-temperature system, gut-associated microbial ecosystems and a fish community involving multiple species as an exploration of OIF's practical applications. Even though information-theoretic quantifiers have been widely used in the research of complexity science, only a small number of works have been reported on the real-world applications of information-based complex network models, especially complex multivariate systems (multispecies ecosystems, for example).

Chapter 2 introduces the Optimal Information Flow (OIF) model based on information theoretic variables including Shannon entropy, mutual information (MI) and transfer entropy (TE). The efficiency of OIF model in inferring causal interactions is validated by using synthetic data generated by a mathematical model, as well as real-world data sets of a sardine-anchovy-temperature system and a multispecies fish community in Maizuru Bay, and by comparing the results from these data sets to the well-documented CCM model that is used here as benchmarks. The results of this study show that the proposed OIF model not only presents a good performance in casual inference especially for highly nonlinear and weak coupling, but provides a broad ecological information by extracting predictive species interaction networks from time-series data.

Chapter 3 uses the OIF model with an improvement to tackle the complexity of gut-associated microbial ecosystems for the Irritable Bowel Syndrome (IBS) that is one of the most prevalent functional gasintrointestinal disorders in human populations. An optimal threshold for TE-based interactions that maximizes the information content is used to reconstruct species interaction networks. This novel complex network model is able to identify the difference in structure and function between healthy and unhealthy networks. Macroecological analyses are also performed for species abundance and diversity indicators (α, β, γ-diversity). Through linking these results to the OIF-inferred species interaction networks, it is observed that the magnitude of species interactions is related to species abundance. These

species-specific results are useful for public health and disease diagnosis.

Chapter 4 explores the application of OIF model in a multi-species marine fish ecosystem. Considering a multi-model comparison, Kernel estimator is selected as the estimator for TE computation in this study. The objective of this work is to investigate how the fluctuations of sea temperature affect the system dynamics and collective behavior of the fish community, and to recognize the critical species that are most responsible for the system stability and sustainability by analyzing the fish ecosystem at the level of both system and species on temporal and temperature-dependent scales. To this end, the time series recording the observations of fish species are categorized into five temperature ranges: $\leq$10°C, 10-15°C, 15-20°C, 20-25°C, $\geq$25°C. Macroecological analysis and OIF network modeling are conducted for each temperature range and the whole time series. Differences in macroecological indicators and functional networks among temperature ranges show particular features of biomass and taxonomic diversity, and system dynamics and stability, implying the impacts of sea temperature on the fish community. In addition, species-specific analyses also allow to recognize the species that are most affected by the change of sea temperature.

The combined results of these three studies show both methodological and application-specific insights related to information dynamics of ecosystems. In fact, ecosystems themselves are highly complex and dynamical systems that demand to be studied not only at individual level, but at system level by characterizing information dynamics among all individual components and treating them as integrated systems on multiple scales. The proposed OIF model and OIF-inferred dynamical networks in conjunction with macroecological analysis allow to study these issues simultaneously and investigate the dynamical evolution of ecosystems under the pressure of external stressors. The results of this research would help to improve the resilience of ecosystems to various disturbances, making ecosystems healthier and more sustainable.

Chapter 2

Inferring Ecosystem Networks as Information Flows

2.1 Introduction

2.1.1 Ecosystem Complexity and Predictability

The flourishing development of complexity science [32, 33] has shed light on research questions and applications in many interdisciplinary fields, for instance, climate change [34–36], epidemiology [37, 38] and ecosystem sciences at multiple scales [39, 40]. In this burgeoning science, complex network models play a central role in the quantitative analysis and design of ecosystems and their representation. This is because functional and structural networks – such as species interactions and habitat corridors – are the core elements of ecosystems defining species organization. When inferring networks, causal inference [41] is one of the fundamental steps for ecosystem reconstruction and graphical representation by assessing interactions or interdependencies – between biota, environment, and among those – that can be thought as information fluxes in more general terms [42].

In a quantitative sense, network inference can be performed via causal inference based on time-series data defining the dynamics of ecosystem components. Causal inference also attracts much attention in some emerging disciplines such as big data science via machine learning since it brings a new set of tools and perspectives for some problems in these areas. However, this issue of causal inference is still an extremely challenging problem due to the intrinsic lack of knowledge or observability of the ”true” reality of a system especially for highly complex non-linear systems

driven by non-linear environmental forcing. Certainly the objective of causal inference is defining unknowns; however robust model validation must be performed. In order to make causal inference practical and achievable, causality is often replaced with predictability as it is articulated in this paper. A plethora of conceptual approaches, frameworks and algorithmic tools including but not limited to Pearson's Correlation Coefficient (PCC) [43, 44], Bayesian Networks (BNs) and Dynamic Bayesian Networks (DBNs) [45–49], Neural Networks, Graphical Gaussian Models (GGMs) [50, 51], Wiener-Granger causality (GC) model [19], Structural Equation Modeling (SEM) [52–57], Convergent Cross Mapping (CCM) [20] and information-theoretic models [58–61] for instance, to tackle causal interactions and infer complex networks in terms of correlation, predictability and probability have been well established; however, most tools are solely tested on low-dimensional systems and some are even untested on ecosystems at different levels of complexity or simulated ones. The vast majority of these models in ecosystem science (with the exception of CCM and few others such as PCMCI [62]) consider only the inferred causality between species pairs one at at time without the simultaneous consideration of all species pairs for each species that is shaping ecosystem collective behavior mediated by environmental dynamics. The ensemble of all species causations is representable as nonlinear dynamical network over the space-time-environmental domain considered. It is therefore valuable for science to seek for robust models and explore novel methods to identify and quantify the pattern-oriented causality between variables (such as species) and how this causality is predictive of target complex system patterns.

2.1.2 Optimal Information Flow Model

Causal inference models and their pros and cons are reviewed in Chapter 1. In this study, considering the limitations of CCM, information-theoretic TE is used to infer causal interactions in a predictive sense and improved considering its extension over time to reconstruct dynamical information networks, the varied Markov order of each variable and a more refined pattern-oriented criteria to select optimal threshold based on maximization of Mutual Information, forming the integrated Optimal Information Flow (OIF) model. The proposed OIF overcomes the limited TE for the directed uncertainty reduction scheme (already present in Li and Convertino (2019) [31]), for the consideration of maximum information/entropy via considering the full network entropy pattern (as in Servadio and Convertino (2018) [30]),

and the MI-based maximization criteria to define the interaction threshold to accurately predict system's patterns (e.g. biodiversity).

The performance of OIF is assessed by applying it to three prototypical case studies including mathematical deterministic and real-world ecosystems. One is a biologically inspired mathematical model that can generate synthetic two-coupled time-series variables describing dynamics similar to predator-prey dynamics [20]. Two parameters (β_{xy} and β_{yx}) in the equations underlying the model are describing the strength of true interdependence between two simulated variables and they can be free varied. Other two case studies are real-world ecosystems: the case of externally forced poorly coupled species (sardine-anchovy-temperature system) [20] and the one of highly complex interacting species (fish community in Maizuru bay) [21] (Figure 1). The well-documented CCM method is also used for these three cases and the results from CCM, despite its known drawbacks of convergence, possible asymmetrical causality miscalculation, and computational complexity, are somewhat considered as benchmark interactions due to the lack of other estimates.

OIF is not perceived as a competitor with other already published models, including GC and CCM, but rather it aims at providing an alternative and hopefully more precise assessment to predictive inference in cases not completely covered by previous models. Theoretically, leaving aside systematic data issues, OIF is expected to give a better performance than other models in interdependence assessment owing to the aforementioned fine properties of TE for nonlinear dynamics. Besides, given the relationship between entropy and diversity [63] (specifically Shannon and Transfer entropy and α- and β-diversity), OIF provides a potential advantage to predict the information about macroecological indicators of ecosystems. In consideration of these features, TE causality is proposed as *non-linear predictability* of both population abundance of species and community macroecological indicators, simultaneously.

2.2 Methods

2.2.1 Ecosystems Models

Bio-inspired Two Species Mathematical Model

In [20], a mathematical model was introduced to generate coupled nonlinear sequences for testing the CCM presented in that study. The model consists of two

diffusively coupled logistic maps describing a simple bio-inspired dynamics without any external environmental effects on both species. It is analytically formulated as:

$$\begin{aligned} X(t+1) &= X(t)[r_x - r_x X(t) - \beta_{xy} Y(t)] \\ Y(t+1) &= Y(t)[r_y - r_y Y(t) - \beta_{yx} X(t)] \end{aligned} \tag{2.1}$$

where X and Y are two random variables linked by factors β_{xy} and β_{yx} that establish the strength of their interactions. It gives possibility to estimate the "true" causality in an absolutely numerical sense and this model can be therefore indicated as $S(\beta_{xy}, \beta_{yx})$. If β_{xy} is fixed as 0 and β_{yx} is fixed as non-zero or varied free, that is, X causes Y, but not vice versa. If β_{xy} and β_{yx} are both non-zero, X causes Y and vice versa. These conditions generate two different kinds of coupling variables that respectively represent unidirectionally and bidirectionally interactive species-species systems described as the case 1 in Figure 2.1. r_x and r_y are the intrinsic growth rates for each variable.

In this study, we focus on both unidirectionally and bidirectionally interactive species-species systems. For unidirectional coupling, β_{xy} is fixed as 0, but β_{yx} is free varied within the range of [0,1], leading to a simplified model as $S(0, \beta_{yx})$. This unidirectional model $S(0, \beta_{yx})$ is exploited to generate coupled sequences of two random variables X and Y where X affects Y, but not vice versa. For bidirectional coupling, we study two different conditions. On the one hand, we consider a system $S(\beta_{xy}, \beta_{yx})$ in which β_{xy} is fixed as 0.2, 0.5 and 0.8, β_{xy} is free varied within the range of [0,1]. This model can be indicated as a univariate model $S(0.2/0.5/0.8, \beta_{yx})$ and is used to generate interdependent coupled time-series variables where the effect of X on Y changes over β_{yx}, while the effect of Y on X is stablized due to the fixed non-zero β_{xy}s. On the other hand, we consider a bivariate system $S(\beta_{xy}, \beta_{yx})$ in which both β_{xy} and β_{yx} are free varied. This model indicated as $S(\beta_{xy}, \beta_{yx})$ generates coupled time series of variables X and Y where X and Y randomly interact with each other. For all these conditions mentioned here, $S(0, \beta_{yx})$, $S(0.2/0.5/0.8, \beta_{yx})$ and $S(\beta_{xy}, \beta_{yx})$ are run under the conditions of which initial $x(1) = 0.4$, $y(1) = 0.2$ and intrinsic growth rates of variables r_X and r_Y respectively are 3.8, 3.5 as used in [20]. These time series of variables X and Y generated by the models for example can be time-series data of species abundance.

Sardine-Anchovy-Temperature Ecosystem

As a real-world ecosystem case study, yearly time-series data of Pacific sardine landings, northern anchovy landings and sea surface temperature (SST) obtained at Scripps Pier and Newport Pier, California are used here. Sardines and anchovies seldom interact with each other because of geographical distribution. External environmental factors including SST affect sardines and anchovies, but not vice versa. It is a typical example of unidirectional causal relationship in real-world ecosystems and such tripartite relationship can be described as the case 2 in Figure 2.1. Sugihara et al. (2012) [20] also studied this fishery ecosystem and successfully inferred the weak causal interactions between between sardines, anchovies and SST with the CCM method. We remake the experiments Sugihara et. al did, and apply our proposed OIF model to do the same work as well, thereby validate the OIF model by comparing results to those from CCM.

Fish Community in the Maizuru Bay Ecosystem

Long-term time-series data counting the observations of the fish community collected along the coast of the Maizuru Fishery Research Station of Kyoto University [21] are used in the multispecies case study described as the case 3 in Figure 2.1 for OIF model validation. Underwater direct visual censuses were conducted approximately once every two weeks from January 1, 2002 to April 2, 2014, totally generating 285 time points sequences during about 12 years long census. The total number of species observed here was greater than 1000, while most species were rare and were not observed during most census time points. If so, the dataset may include a lot of zero values making it difficult to process. Therefore, only 14 dominant fish species and 1 jellyfish species were selected in this dataset.1 Jellyfish species was selected in the dataset because this species was abundant in this area and was thought to have considerable influences on the community and ecosystem dynamics. Accordingly, both OIF and CCM model are exploited to measure causal interactions among 14 dominant fish species and 1 Jellyfish setting up the dynamical complex multispecies system.

2.2.2 Interactions Inference Models

Linear Correlation Model

The linear correlation between non-lagged random variables X and Y is given by:

$$\text{corr}(X,Y) = \frac{\sum_{t=1}^{L}(x_t - \bar{X})(y_t - \bar{Y})}{\sqrt{\sum_{t=1}^{L}(x_t - \bar{X})^2 \sum_{t=1}^{L}(y_t - \bar{Y})^2}} \tag{2.2}$$

where $\bar{X} = \frac{1}{L}\sum_{t=1}^{L} x_t$ and $\bar{Y} = \frac{1}{L}\sum_{t=1}^{L} y_t$. L is the length of time-series of X and Y.

Convergent Cross Mapping Model

The principle of CCM model involves state space reconstruction from two variables and quantifies the potentially causal (asymmetrical) relationship between these variables using the method of nearest neighbor forecasting. Nearest neighbor forecasting method is an application of Takens' Theorem called simplex projection. States of a system are reconstructed by applying successive time lags of time-series variable underlying the method of time lag embedding [20, 22]. Interestingly, this method has been originally applied to describe the transition to turbulence of fluids [64, 65].

In the case where X causes Y, Takens' theorem indicates that there should exist a "causal" relationship between states of X and the contemporaneous states of Y. CCM quantifies this relationship using the simplex projection to predict time-series X from reconstructed Y. Specifically, a manifold M_X ("reconstructed", "shadow" or predictor manifold) is constructed from lags of variable X (i.e., $X(t-\tau)$ with the time lag τ) and used to estimate contemporaneous values of $Y(t)$. M_X is an approximation that will display convergence up to the level set by observational error and process noise. At convergence the approximated $\hat{Y}(t)|M_X$ will be close to $Y(t)$. The relationship between $Y(t)$ and $Y(t-\tau)$ is on the target manifold. To explore the opposite "causality" CCM explores the convergence of $\hat{X}(t)|M_Y$ to $X(t)$ where M_Y is the predictor manifold. Thus, CCM determines how well local neighborhoods (defined by $E+1$ points, that is the minimum number of points needed for a bounding simplex in an E-dimensional space) on the manifold M_X correspond to local neighborhoods on M_Y. The R package "rEDM" is available

online https://rdrr.io/cran/rEDM/ through the Comprehensive R Archive Network (CRAN).

Pearson's correlation coefficients ρ (originally defined as Eq. 2.2 considering non-time lagged variables) between predicted time series and observations of X (or Y) are calculated. The non-linear correlation coefficient is considered as the indicator of cross-mapping skill, that is the "causality" between species X and Y. The non-linear ρ is defined as:

$$\begin{aligned} \rho(X \to Y) &= \frac{\text{cov}(Y(t), \hat{Y}(t)|M_X)}{\sigma_Y \ \sigma_{\hat{Y}(t)|M_X}} \\ \rho(Y \to X) &= \frac{\text{cov}(X(t), \hat{X}(t)|M_Y)}{\sigma_X \ \sigma_{\hat{X}(t)|M_Y}} \end{aligned} \tag{2.3}$$

where cov and σ are the covariance and standard deviation. $\hat{X}(t)|M_Y$ and $\hat{Y}(t)|M_X$ are the predicted values of $X(t)$ and $Y(t)$ considering the attractor manifolds of lagged Y and X. Considering the relationship between the calculated cross-mapping skill and the length of time series L, ρ increases with L until a convergent stable value. ρ is alway larger the longer L and that indicates causality according to [20]. Typically, no less than 30 points in the time-series data should be used for CCM analyses [66]. Further details about the use of CCM for this study are provided in Supplementary Information.

Optimal Information Flow Model

Transfer entropy (TE) [23] is a non-parametric statistic in information theory that estimates the amount of information that a source variable contains about a destination variable considering destination's current and historical states. It measures how much directed (time-asymmetric) information transfers between two variables, giving an incentive to quantify the causal relationship between two variables with TE. Here it can be calculated as:

$$TE_{Y\to X}^{(k,l,u)} = \sum_{x,y} p(X_{t+1}, X_t^{(k)}, Y_{t-u+1}^{(l)}) log \frac{p(X_{t+1}|X_t^{(k)}, Y_{t-u+1}^{(l)})}{p(X_{t+1}|X_t^{(k)})} \ , \tag{2.4}$$

Where, X and Y denote two random variables, k and l refer to the Markov orders of variables X and Y, implying that we need to at least consider k (l) time points of variable X (Y) in the past for the estimation in order to capture all relevant

information in the past of X (Y). Here we assume that the time-series analysis here obeys a memoryless Markov process. Hence, parameters k and l are fixed as 1, meaning that the next states of X and Y are only dependent on the current states and not on states in the past. u is the source-target time delay establishing lagged influence and is freely varied. Servadio and Convertino (2018) [30] proposed a framework of optimal information networks (OIN) to select TEs that maximize the total entropy for inferred networks. Probability distribution functions predicted by the network with maximum entropy are considered as the most general distributions fitting the observations.

In this study we also investigate the optimal TE that provides a clearer detection and more accurate quantification for causal relationships between two variables by choosing an appropriate time delay u in a specified range. The choice of the optimal u within a range leads to the optimal TE model and resultant network inference while considering the minimum computational complexity. As shown in our previous work about microbiome [31], time delay u used to calculate TE between species is the one who minimizes the distance from one species to another in the inferred network.

The distance can be calculated by [59]:

$$d(X, Y) = e^{-MI(X(t\pm u);Y(t))} \tag{2.5}$$

where MI is the mutual information of variables X and Y. MI of two random variables X and Y, is given by:

$$MI(X(t \pm u); Y) = \sum \sum p(x(t \pm u), y)\ log \frac{p(x(t \pm u), y)}{p(x)p(y)}\ , \tag{2.6}$$

where $p(x)$ and $p(y)$ are the marginal distributions of random X and Y, and $p(x(t \pm u), y)$ is the joint probability distribution that is the pdf affected by the time-delay. The time delay u that is chosen is exactly the one that maximizes the MI of the two variables, because that is the one that minimizes the uncertainty. MIs is calculated using a range of time delays u (defining temporal entropy reduction parameters) and then the time delay corresponding to the maximum MI is selected before the calculation of TE. The choice of using the time delay u that maximizes MI is focusing on the highest predictability rather than the investigation of true causality. For example, two species may have a relatively low interaction except for a limited time period when the interaction is very high. Thus, by considering the maximum MI the focus is on the highest predictability vs. the average most likely

mutual dynamics, or more precisely it is focused on the magnitude of potential interaction rather than the frequency. Yet, extreme accidental interactions are also captured by the choice of $MI(u = u_{max})$, although very small or non-existent interactions exist.

In addition, we also compare the results of the linear cross-correlation estimates and the non-linear MI estimates for species interactions. Mutual information is a distance between two probability distributions while correlation is a linear distance between two random variables. This is done to detect the causal relationships between variables in the non-linear mathematical predator-prey model (where the two variable X and Y can be also belong to the same species) also to detect the performance of linear vs. non-linear interaction inference models. On the contrary of the asymmetric TE (that measures directed interdependencies between variables), MI, as well as cross-correlation, provides a symmetric measure for inferring mutual interdependencies unable to identify the direction of potential causal interactions.

2.2.3 Predicted Ecosystem Biodiversity Patterns

Taxonomic and Effective α-diversity

We consider the macroecological indicator α-diversity (or taxonomic diversity) for the fish community in the Mairuzu Bay (Kyoto, Japan) to investigate whether the OIF model infers a "causal" network able to predict α-diversity over time with high accuracy. Results are compared to the α-diversity calculated from the CCM inferred network. The OIF and CCF α-diversity are introduced as "effective" α-diversity to consider the estimated interacting species rather than counting species independently of the interaction. Specifically, we base our analysis on taxonomic α-diversity that is the most elementary definition of local biodiversity of a community. However, taxonomic α captures only one aspect of diversity that may not be sufficient especially for very uneven communities where species have very different abundance. Nonetheless, this does not affect our model intercomparison, since any model discussed here can be used for any diversity metric such as Shannon index and Simpson diversity.

α-diversity is a concept in ecology that counts the number of species (biodiversity) observed at a local scale in space and time. In this multispecies case study, the local scale is a time-dependent measure for the whole community. The resolution at which α is assessed (i.e., the sampling interval of the time-series data) is two weeks.

For a set of species $\mathbf{S} = \{S_1, S_2, ..., S_n\}$ whose abundance $\mathbf{X} = \{X_1, X_2, ..., X_n\}$ changes over time, $\alpha(t)$ can be calculated as:

$$\alpha(t) = \sum_{k=1}^{n} x_k(t)^0 \ , \tag{2.7}$$

where $x_k(t)$ is the abundance of species k at time point t.

An effective α-diversity is also derived from the inferred species interaction networks using CCM and OIF models. The estimated effective α-diversity (indicated as $\alpha_E(t)$, hereafter) is the number of nodes (species) in the inferred networks considering the minimum data length l required for the inference. The estimated α-diversity from CCM and OIF can be obtained as:

$$\alpha_E(g) = \sum_{i=1}^{n} k_i(g) \tag{2.8}$$

where,

$$k_i(g) = \begin{cases} 0, & \text{for } \sum_{j=1}^{n}(|M_{i,j}(g)| + |M_{j,i}(g)|) = 0 \\ 1, & \text{for } \sum_{j=1}^{n}(|M_{i,j}(g)| + |M_{j,i}(g)|) \neq 0 \end{cases} \ , \tag{2.9}$$

denotes the structural degree of all nodes (species) involved in the inferred networks for a time period $g = t - l$. $M(t)$ is the $n \times n$ interaction matrix from OIF or CCM model for each time period g.

The total number of time periods on which the inference of networks is performed (and $\alpha_E(g)$ is calculated) by both OIF and CCM is $G = \lfloor \frac{L-l}{\Delta t} \rfloor + 1$. L is the total number of observations of abundance obtained every two weeks for each species, l is the minimum number of observations set up to perform the inference of interactions (i.e., l=30 in this study supported by the evidence about the minimum length required to perform a robust inference of ρ with CCM) for the whole time series L, and Δt is the numerical inter-observation time (or time step) that corresponds to two weeks. $\lfloor \bullet \rfloor$ (where $\bullet = \frac{L-l}{\Delta t}$) rounds G to the smaller integer. Note that l is the maximum embedding dimension E for CCM. In this multispecies case study, the length of raw time series is 285 and the time step is chosen as 1. Thus, G is equal to 256. CCM and OIF models leverage 256 shortened time series to estimate the potential causality between all possible pairs of fish species, leading to 256 dynamical networks. Note that the number of $\alpha(t)$ values is higher than the number of $\alpha_E(g)$ because of the need of a minimum data length of network inference models to infer species diversity.

Simpson's Diversity and Shannon Indices

The Simpson's Diversity Index (SDI) has been calculated for comparison with α-diversity. SDI is a macroecological indicator that gauges diversity differences in communities considering also species population abundance. The calculation of SDI is given by:

$$SDI(t) = 1 - \frac{\sum_{k=1}^{n}[x_k(t)(x_k(t)-1)]}{\sum_{k=1}^{n} x_k(t)\ (\sum_{k=1}^{n} x_k(t) - 1)};. \tag{2.10}$$

The range of SDI is from 0 to 1 (1 is the total normalized diversity) in which high scores indicate high diversity and low scores indicate low diversity.

In light of the relationship between entropy and diversity introduced by [63], we also investigated the Shannon index that is defined as the sum of entropies of all species based on time-series abundance. This is the first component of the information balance equation as introduced in [31]. The Shannon index is formulated as:

$$H_\alpha(t) = \sum_k H(x_k(t)) = -\sum_k p_k(t)\ \log\ p_k(t) \tag{2.11}$$

where $p_k(t)$ is the probability to observe species k at time point t; more specifically this probability is based on the abundance of each species at each time step. Probabilities are obtained from probability density function estimates via kernel density estimation (KDE). $H_\alpha(t)$ gives the uncertainty as diversity information index rather than the taxonomic diversity and describes how species are assembled together via their probabilistic nature. Distributions reflect dynamics of species instead of considering simple occurrence as a binary variable. We compare $H_\alpha(t)$ to taxonomic $\alpha(t)$ and SDI to recognize similarities and differences in biodiversity indicators (or patterns more generally). This emphasizes the differential sensitivity of interactions' importance for different indicators, whether one considers $\alpha(t)$ or SDI for instance. Different indicators, or patterns, reveal different information about the ecosystem analyzed.

2.3 Results

2.3.1 Two Species Unidirectional Coupling Ecosystem

This bio-inspired ecosystem S($\beta_{xy} = 0$, β_{yx}) describing the unidirectional coupling is run for 1000 time steps for reaching stationarity, generating a set of 1000

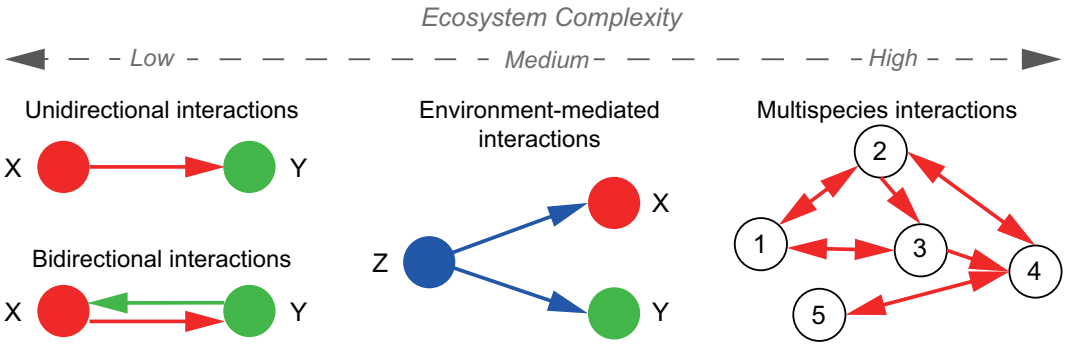

Figure 2.1: **Studied ecosystem complexity.** Epitomes of increasing ecosystem complexity are shown from left to right where nodes are representing variables (e.g. species or other socio-environmental features). Case 1 shows two basic cases: unidirectional and bidirectional interactions where true interaction strength is known because embedded into a mathematical model. Case 2 is about environment-mediated interactions with no knowledge of "true" interactions. Case 3 is a multispecies ecosystem with multiple bidirectional interactions with no knowledge of "true" interactions.

points long time-series dependent on β_{yx}. This means that species X has an increasing effect on Y with the increase of β_{yx}, but Y has no effect on X. Both CCM and the proposed OIF model are separately used to quantify the potential causality between species X and Y. The inferred causality dependent on β_{yx} (as physical interaction) only is shown in Figure 2.2A. Figure 2.2A shows that under the condition of β_{xy}=0, results of "Y to X" (i.e. the estimated effect on Y on X) is close to 0 for the OIF model ($TE_{Y\to X}(\beta_{yx})$) that precisely describe the no-effect of Y on X. "X to Y" ($TE_{X\to Y}(\beta_{yx})$) well tracks the increasing strength of the effect of X on Y for increasing values of the physical interaction β_{yx} embedded into the mathematical model. However, considering results of the CCM model, "Y to X" ($\rho_{Y\to X}(\beta_{yx})$) presents non obvious (and likely wrong) non-zero values with higher fluctuations compared to $TE_{Y\to X}(\beta_{yx})$ especially for lower values of β_{yx}. This erroneous estimates of CCM is likely related to the need of CCM for convergence. For CCM, "X to Y" (($\rho_{X\to Y}(\beta_{yx})$)) shows an increasing trend for increasing values of β_{yx} and decreasing when β_{yx} is greater than $\sim$0.5 non-trivially. In consideration of these results for the unidirectional coupling ecosystem, the OIF model performs better over CCM in terms of unidirectional causality inference.

2.3.2 Two Species Bidirectional Coupling Ecosystem

In this case, the effect between two species is bidirectional. Species X has an effect on species Y and vice versa. The univariate dynamical systems S(0.2/0.5/0.8, β_{yx}) are run for 1000 time steps under the same conditions determined by β_{xy}. Certainly this situation is fictional since in real ecosystems the interaction strength is changing when other interacting species change their interactions.Thus, keeping one interaction fixed around one value is a strong unrealistic simplification (analogous of one-factor at-a-time sensitivity analyses) but it is a toy model that allows to verify the power of network inference models. These models generate three sets of 1000 points long time-series dependent of β_{yx} for each fixed β_{xy}. OIF and CCM are used to infer "causality" between X and Y – in the form of ρ and TE – and compare that against the real embedded interaction β_{yx} and β_{xy} shown in Figure 2.2BCD. Considering all results of Figure 2.2 corresponding to fixed β_{xy}s, the correlation coefficient ρ yielded from CCM and TE from OIF are both able to track the strength of causal trajectories. However, TE seems to perform better in term of ability to infer fine-scale changes in interactions. In particular, considering Figure 2.2D (right plot), higher TE_{yx} higher for low β_{yx} makes sense because $\beta_{xy} > \beta_{yx}$ that means Y has a larger influence on X than vice versa and then Y is able to predict X. Additionally, TE does not suffer of convergence problems; specifically, considering Figure 2.2A (left plot), higher ρ for small β_{yx} is not sensical and that is likely related to convergence problems of CCM.

Additionally, $\rho_{Y\to X}(\beta_{yx})$ shows higher fluctuations on average especially for the condition of lower β_{yx}s compared to $TE_{Y\to X}(\beta_{yx})$. When considering the effect of X on Y that is a function of β_{yx} for CCM, $\rho_{X\to Y}$ reaches an extreme value at around β_{yx} = 0.5 and then declines for larger values of β_{yx}. This is not consistent with the expected effect of X on Y that should be proportional to β_{yx} embedded into the mathematical model. The ability of ρ to reflect the proportional relationship between the effect of X on Y (manifested by β_{yx}) vanishes for high β_{xy}s due to unexpected and somewhat inconspicuous changes in $\rho_{X\to Y}$ for larger β_{yx}. In simple words, the expected increasing trend of ρ is lost for larger β_{xy} that is counterintuitive. On the other side, $TE_{X\to Y}(\beta_{yx})$ invariably maintains an increasing trend for increasing values of β_{xy}. OIF is also performing better than CCM when predicting higher average values of $TE_{Y\to X}$ for increasing values of β_{xy} (red curves in Figure 2.2ABCD, right plots) as expected by the fixed effect in the mathematical model of Y on X. These results suggest that when compared to ρ of CCM, TE can track well

the causal interactions over β_{yx} with higher performance and without considering the convergence requirement of CCM. CCM needs to consider the length of time series that makes $\rho_{X\to Y}(\beta_{yx})$ convergent to a stable value, but uncertain for large differences in time-series length of (X,Y) and sensitive to short time series.

In more realistic settings for real ecosystems (and in analogy to global sensitivity analyses) when β_{xy} and β_{yx} are both considered as arguments of the two-variable (X,Y) bio-inspired model, the simulated ecosystem becomes a truly bivariate system, yet yielding complexity but more interest into the causality inference (Figure 2.3). The dynamical system S(β_{xy}, β_{yx}) was generated for 800 time steps under the same conditions mentioned above. We generated the datasets that allowed us to study linear and non-linear predictability indicators for inferring the embedded physical interactions. Specifically, we measure undirected linear correlation coefficient $corr_{X;Y}(\beta_{xy}, \beta_{yx})$, non-linear undirected mutual information $MI_{X;Y}(\beta_{xy}, \beta_{yx})$, directed non-linear correlation coefficient $\rho_{X\to Y}(\beta_{xy}, \beta_{yx})$ and $\rho_{Y\to X}(\beta_{xy}, \beta_{yx})$, and non-linear directed transfer entropy $TE_{X\to Y}(\beta_{xy}, \beta_{yx})$ and $TE_{Y\to X}(\beta_{xy}, \beta_{yx})$ as shown in Figure 2.3.

These 2D phase-space maps in Figure 2.3 show strikingly similar patterns for classical linear correlation coefficients, MI, ρ of CCM and TE of OIF which underline the fact that all methods are able to infer the interdependence patterns of interacting variables explicitly defined by β_{xy} and β_{yx}. The color of phase-space maps is proportional to the inferred interaction between X and Y when the mutual physical interactions are varying according to the mathematical model in Eq. 2.1.

In Figure 2.3, even though phase-space maps of undirected $corr_{X;Y}(\beta_{xy}, \beta_{yx})$ and $MI_{X;Y}(\beta_{xy}, \beta_{yx})$ present similar patterns (in value organization and not value range) to those of directed ρ and TE, neither $corr_{X;Y}((\beta_{xy}, \beta_{yx})$ and $MI_{X;Y}((\beta_{xy}, \beta_{yx})$ provide information about the direction of causality. As expected MI shows the opposite pattern of the average TE due to the fact that MI is the amount of shared information (or similarity) versus the amount of divergent information (divergence and asynchronicity) between X and Y. In a biological sense TE should be interpreted as the probability of likely uncooperative dynamics (leading to or driven by heterogeneity) while MI as the probability of cooperative dynamics (leading to or driven by homogeneity). Certainly, cooperation in a biological sense should be interpreted on a case by case basis. In a broader uncertainty propagation perspective [30], "cooperation" between variables means that variables contribute similarly to the uncertainty propagation, while "competition" means that one variable is predominant over the other in terms of magnitude of effects since TE is proportional

to the magnitude rather than the frequency of effects. For the former case the total entropy of the system is higher than the latter case. Interestingly, correlation corr(β), ρ and TE show similar patterns in both organization and value range (but not in singular values of course), which sheds some important conclusions about the similarity and divergence of these methods as well as their capacity and limitations in characterizing non-linear systems.

When comparing the phase-space patterns from CCM and OIF (displaying ρ and TE) a more colorful and informative pattern is revealed by OIF. This means that TE gives a better gradient when tracking the increasing strength of causality for increasing values of β_{xy} and β_{yx}. When comparing the phase-space patterns for the two causal directions of "$X \rightarrow Y$" and "$Y \rightarrow X$", phase-space maps from CCM are almost similar, while those from TE present apparent differences in the strength of effects for these two opposite directions. Therefore, TE model is much more sensitive to the direction compared to CCM when detecting the directional causality. These results imply that TE performs better to distinguish different embedded physical interactions (dependent on direct interactions β-s, but also growth rate r_x and r_y, and contingent values $X(t)$ and $Y(t)$ determining the total interaction as seen in the model of Eq. 2.1) in the causal relationship between species. It should be emphasized how all linear and non-linear interaction indicators are inferring the total interaction and not only those exerted by β-s. In a broad uncertainty purview [30] the importance of these three factors depends on their values that define the dynamics of the system; dynamics such as defined by the regions identified by patterns in Figure 2.3 for the predator-prey system in Eq. 2.1. Figure 2.4 highlights three different dynamics corresponding to the TE blue, green and red regions in Figure 2.3.

In all dynamical states represented by Figure 2.3, species are interacting with different magnitude and this defines distinct network topologies. Three prototypical dynamics are show in Figure 2.4 with colors representative of ρ and TE in Figure 2.3. The "blue" deterministic dynamics has very high synchronicity and no divergence considering variable fluctuation range (the gap is deterministic and related to the numerically imposed $u = 1$), as well as no linear correlation between non-lagged variables. In perfect synchrony one would have one point in the phase-space. Thus, absence of correlation does not imply complete decoupling of species but it can be a sign of small interactions. The "green" dynamics shows a relatively high synchronicity and medium divergence. In the phase-space of synchronous values of X and Y a correlation is observed with relatively small fluctuations because

the divergence is small. Lastly, the "red" dynamics shows a relatively high asynchronicity and divergence. The stochasticity is higher than previous dynamics and the "mirage correlation" in the phase space has higher variance. Mirage correlations means that correlation does not imply similarity in dynamics for the two species. Non-linearity is higher from blue to red dynamics as well as predictability but lower absolute information entropy. Then, it is safe to say that linear dynamics (or small stochasticity) does not imply higher predictability.

2.3.3 Real-world Sardine-Anchovy-Temperature Ecosystem

CCM and proposed OIF model are also used for a real-world fishery ecosystem to infer potential causal interactions between Pacific sardines (*Sardinops sagax*) landings, Northern anchovies (*Engraulis mordax*) and Sea-Surface Temperature (SST) recorded at Scripps Pier and Newport Pier, California. Sardines and anchovies do not interact physically (or the interaction is low in number), while both of them are influenced by the external environmental SST that is the external forcing. To quantify the likely causal interactions between species and SST based on real data, we use CCM considering the length of time series for convergence of ρ, as well as OIF considering a set of time delays for acquiring stable values of inferred interactions TEs.

Results from CCM in Figure 2.5A (plots from top to bottom) show that no significant interaction can be claimed between sardines and anchovies, as well as from sardines or anchovies in the SST manifold which expectedly indicates that neither sardines nor anchovies affect SST. This latter results, considering its biological plausibility should be taken as one validation criteria of predictive models, or complimentary as a test for anomaly detection of spurious interactions. The reverse effect of SST on sardines and anchovies can be quantitatively detected with the correlation coefficient ρ as well as TE. Although the calculated causations between SST and sardines or anchovies are moderate, CCM is able to provide a good performance in causality inference when the length of time series used is long enough due to convergence requirement.

Figure 5B shows OIF's results of inferred causal interactions between sardines, anchovies and SST dependent on the time delay u. For sardines and anchovies, OIF exposes the elusive bidirectional interactions that is actually biologically plausible versus the results of CCM that infer $\rho = 0$. For the effect of external SST on sardines and anchovies, OIF model gives unstable causal interactions with bias for

lower time delays due to known dependencies of TE on u (such as cross-correlation for instance) that establishes the temporal lag on which the dependency between X and Y is evaluated. In a sense, plots in Figure 5B are like cross-variograms for the pairs of variables considered. TE becomes stable when the time delay is located in an appropriate range. It means that OIF requires an optimal time delay that makes results of the causality inference robust and that is related to the optimal transfer entropy model (as highlighted in [31] and [30]) that defines the most likely interdependency between variables. As much as ρsensitivity of TE is also observed for small time series that do not allow to infer probability distribution functions sufficiently well. However, the length of data L is a factor affecting ρ more than TE. The finding from OIF and CCM that SST unidirectionally affects the size of the population of sardine and anchovy corroborate earlier findings of [67] who detected the correlation between 3-year running average recruitment and spawning stock size, as well as the previous results from CCM [21].

Figure A.1 shows the relationships between normalized ρ and TE estimated for all selected values of L and u of pairs in Figure 2.5 (sardine-anchovy, sardine and SST, anchovy and SST). These plots show opposite results than the proportionality between ρ and TE in Figure 2.3 because non-optimal values are used, that is non-convergent ρ-s and suboptimal TE during the interaction inference procedure (Figure A.1). TE for too small u-s determines overestimation of interactions due to the implicit assumptions that variables have an immediate effect on each other and that is not alway the case as highlighted by the vast time-lagged determined nonlinear regions in Figure 2.3. If "transitory" values of ρ for small L are disregarded, as well as TEs for small u-s, the relationship between ρ and TE shows a correct linear proportionality.

2.3.4 Real-world Multispecies Ecosystem

Interactions between fish species living in the Maizuru bay are intimately related to external environmental factors of the ecosystem where they live, the number of species living in this region considering also the unreported ones) and biological species interactions, which leads to a complex dynamical nonlinear system. In Figure 2.6 the network of observed fish species (Table A.1) is reported where only the interactions considered in [21] for the CCM are reported. This is because the goal is to compare the CCM inferred network to the TE-based one based on abundance. Figure 2.7 shows the temporal fluctuations of abundance and the functional interac-

tion matrices of ρ and TE. In this paper we study and compare average ecosystem networks for the whole time period considered but dynamical networks can also be extracted via time-fluctuating ρ and TE as show in Figure A.3. These dynamical networks can be useful for studying how diversity is changing over time and ecosystem stability (Figs. S4, S6-S7) as well as understanding the relationship between ρ and TE (Figure A.5). In the network of Figure 2.6 the color and width of links is proportional to the magnitude of TE (Table A.2); for the former a red-blue scale is adopted where the red/blue is for the highest/lowest TEs. The diameter is proportional to the Shannon entropy of the species abundance pdf (Table A.3). The color of nodes is proportional to the structural node degree, i.e. how many species are interconnected to others after. Therefore, the network in Figure 2.6 is focusing on uncooperative species whose divergence and/or asynchronicity (that is a predominant factor in determining TE over divergence) is large. Yet, the connected species are rarely but strongly interacting in magnitude rather than frequently and weakly (i.e., cooperative or similar dynamics). Additionally, the species with the smallest variance in abundance are characterized by the smallest Shannon entropy (smallest nodes) and more power-law distribution although the latter is not a stringent requirement since both pdf shape and abundance range (in particular maximum abundance) play a role in the magnitude of entropy. Average entropy such as average abundance are quantities with limited utility in understanding the dynamics of an ecosystem as well as ecological function. Nonetheless, species with high average abundance (e.g. species 5) have a very regular seasonal oscillations and the largest number of interactions with divergent species. A result that is expected considering the size of the population and the ability of the species to follow regular environmental fluctuations.

Figure S2 shows that the strongest linear correlation is for the most divergent and asynchronous species (from species 4 to 9) for which both ρ and TE are the highest (Figure 2.7 B and C). This confirms the results of Figure 2.3 and the fact that competition (or dynamical diversity more generally) increases predictability. This also highlights the fact that linear correlation among state variables does not imply synchronicity or dynamic similarity as commonly assumed. The interaction matrices in Figure 2.7B and C confirm that TE has the ability to infer a larger gradient of interactions than ρ and the total entropy of the TE matrix is lower than ρ. Pairwise the inferred interaction values by CCM and OIF are different but ρ and TE patterns appear clearly similar and yet proportional to each other.

CCM and OIF models are applied to calculate the potential interactions between

all pairs of species. Figure 2.7B and C show interaction matrices describing the normalized ρ from CCM and TE from OIF model of all pairwise species, respectively. The greater the strength of likely interaction, the warmer the color. These results demonstrate that CCM and OIF model present similar patterns for the interaction matrices in terms of interaction distribution, gradient and magnitude in order of similarity. This indicates that both CCM and OIF are able to infer the potentially causal relationships between species. Precisely, the most interacting species (4-9) are the most divergent and asynchronous species as well as diverse in terms of values of abundance; these species form the "collective core" that is likely determining the stability of the ecosystem. When considering their abundance values at the same time steps (Figure A.2) these species are linearly related and this increases their mutual predictability by either using linear or non-linear models based on correlation coefficient and TE. The choice of the optimal u that maximizes MI leads to the optimal TE model and resultant interaction network. The observed u over time is really small (Figure A.9) and this signifies how likely the ecosystem has small memory and responds quickly to rapid changes. The chosen time delay $u = 1$ corresponds to the species sampling of two weeks. Note that values of u are also dependent on the data resolution and they are strongly related to fluctuations rather than absolute biodiversity value. Thus, while biodiversity may fluctuate rapidly in time, value of diversity for seasons or longer time periods can have longer memory.

We also study temporally dynamical networks for the fish ecosystem community (see Section 2.3). CCM and OIF model are applied to quantify the causality between all possible pairs of species at each time period by calculating ρ and TE, respectively. Estimated effective α-diversity (Eq. 2.8) from CCM- and TE-based inferred networks at each time point can be obtained and then compared to the taxonomic (or "real") α-diversity. Results are shown in Figure 2.8 and Figure A.6. In the whole time period, the estimated α-diversity from CCM is constant, whereas the global trend of the estimated α-diversity from OIF model slightly decreases over time that is consistent with the global trend of real α-diversity. CCM always predicts a non-zero interaction for all species (including negative values) whereas OIF predicts zero interactions for some species that are then not making part of the estimated effective α-diversity.

Figure 8 shows the effective α diversity from CCM and OIF for an optimal threshold of ρ and TE (i.e., 0.2 and 0.3) that maximizes the correlation coefficient and Mutual Information (MI) between α_{CCM} or α_{TE} and the taxonomic α, respectively. The maximization of the correlation coefficient and MI guarantees that the

estimated effective α are the closest to the taxonomic α. Figure S6 shows effective α for unthresholded interactions and other thresholds. Note that the threshold on TE does not coincide with the value of TE that maximizes the total network entropy (Figure 2.8) and then some of the reported species may not be part of the ecosystem strongly. Thus, this threshold method is also useful to identify species that are truly forming local diversity vs. transient species. Considering the pattern of fluctuations of effective α-diversity from CCM they are poorly unrelated to the real α-diversity, while those from OIF are much more synchronous with seasonal fluctuations of real α-diversity. However, α_{TE} is a bit higher than the average taxonomic α. Both CCM and TE predicts a decrease in α in time that corresponds to an increase in SST. As shown in Fig S6, OIF is attributing higher sensitivity to SST for small interaction species because α fluctuations show seasonality that happens when species follow environmental dynamics closely. Vice versa, CCM is predicting a broader sensitivity for all positively interacting species. These results reveal that OIF gives an effective tool to measure meaningful interdependence relationships between species for constructing temporally dynamical networks where the number of nodes over time (estimated $\alpha(t)$) can reflect closely the taxonomic α-diversity. This allow us to find more reliably how changes of environmental factors (e.g. SST) affect biodiversity in ecosystems. The establishment of thresholds on interactions is also useful for exploring ranges of interdependencies and associated effective α-diversity with respect to the average taxonomic diversity. Supplementary Information contains further elaborations on results.

2.4 Discussion

In the paper the proposed Optimal Information Flow (OIF) model was validated by considering the problem of causality inference of species interactions for ecosystems with different level of complexity and systemic uncertainty: a deterministic mathematical model of predator-prey dynamics, the real Sardine-Anchovy-Temperature triplet ecosystem in the Pacific, and a real multispecies fish ecosystem in Japan. These three case studies are epitomic example of deterministic, low and high complexity dynamics. The mathematical model can be generalized as a model for interaction dynamics between individual or communities of the same species or between two generic variables X and Y. The quantification of interactions was compared to the well-documented CCM model.

Early method of correlation were proved to be neither necessary nor sufficient to estimate the causal relationship between time series variables (mostly due to the fact that any association does not prove causality because models are not surrogate of reality and scale-dependent data are just a sample of ecosystem dynamics), even though it remains a common and heuristic notion [17] [20]. Despite these views we prove the power and limitations of correlation methods with respect to non-linear methods (such as OIF and CCM), dynamics of the complex ecosystem considered, and target patterns to predict that define whether correlation is able to measure interactions. As for the latter, for instance we show how even highly non-linear systems show linearity when non-linear "causal" variables are considered at the same time step (as a virtue of non-linear asynchronicity). Therefore, the scale of analysis considering also the space-time domain with the explicit consideration of lag effects, determines the dynamics that is visible and the model that can be used for predictions.

Granger causality is the primary framework that uses predictability especially for identifying causation, however it is problematic in highly nonlinear systems even with some deterministic states or components. Sugihara et al. (2012) [20] managed to deal with the problem and introduced the CCM model. CCM was well documented and successfully applied to bio-inspired mathematical models, as well as real-world ecosystems [20]. Despite interactions among species or variables, or interdependencies more generally defined, are rarely completely zero and related to patterns of different processes to capture, Sugihara et al. (2012) [20] maintains the view of a deterministic single-value causality. In our opinion calculating causality in an absolute sense between variables is always not only very hard, but also meaningless because the resulting values are dependent on data and models used as well as the predicted patterns for which interactions are calculated for. The very first question should be causality about what? After that the evaluation of the dynamics of the ecosystem coupled to the target patterns to map should drive model selection. Causality is actually predictability of patterns of interest and predictability can be close to true causality for systems with low complexity and noise. The basic principles to interpret predictability are uncertainty reduction and accuracy that can be quantified as the probability of an event to occur given another one (as predictands and predictors, respectively).

From the perspective of information theory that has attracted attention in complex networks research, entropy is the information theoretic description for uncertainty or more precisely lack of organization rather than absolute uncertainty. Un-

certainty is in fact also information about the diversity of values of a complex system (see e.g. [63] that demonstrated how entropies are reasonable indices of diversity) and the distribution of these values determines entropy. The fundamental work of studying complex networks is to untangle complex interdependencies comprising a large number of potential causations between all pairwise nodes (variables), that allows one to predict the collective behavior of complex systems. The intuitive and heuristic notion for this problem in information theory is transfer entropy that measures the uncertainty reduction (or information flow) between nodes (variable). From this conceptual perspective we name the OIF model for inferring potential causality seen as sets of uncertainty reduction networked fluxes. Multiple transfer entropies for one single variable as a function of all others determine non-linearity that cannot be overlooked even when variable interaction is deterministic. Considering entropy as diversity also implies that OIF provides reflections of temporal changes in diversity (e.g. biodiversity) determined by changes in information fluxes.

The bio-inspired mathematical model generates a clean inter-species interaction ecosystem without any noise, that allow us to estimate "true" causality between synthetic species X and Y. The so-called "true" causality means the causation embedded numerically in the parameters in the dynamical equations 2.1 (β_{xy} and β_{yx}). When β_{xy} is fixed as zero only the unidirectional causality ($X \to Y$) exists between species X and Y. Then, any estimator of predictive causality closer to the physical causality β defines the accuracy of the model. Results from Figure 2.2 shows how OIF model outperforms CCM.

Depending on the values of the parameters the model may capture some biological dynamics such as amensalism and commensalism (when β_{xy} or β_{yx} are zero), or predation, competition and mutualism (when both β_{xy} and β_{yx} are different than zero). Biologists define amensalism (i.e. a strong asymmetrical competition) is a type of biological relationship between species in which one species (e.g. X) has a potential negative effect on another (Y), but the second species Y has no detectable effect on the first species X. Biologically speaking, commensalism is another type of biological relationship in which one species (Y) gets benefits while the other one (X) is neither helped or harmed. In a broader complex dynamics perspective it is easier to talk about "cooperation" or "competition" between variables meaning that variables contribute similarly to the uncertainty propagation, or that one variable is predominant over the other in terms of magnitude of effects vs. the frequency.

The generic dynamical characterization allows to avoid pitfalls of the categorical classification of interactions in biology that suffers from the lack of knowledge

about true and meaningful values of interactions that distinguish one biological dynamics from another. We also caution to use numerical estimates of interactions to replace empirical biological knowledge because data-inferred interactions are always much more complex than "lab settings" values and these values are certainly highly affected by the environmental context and measurement technology potentially. The similar pattern of inferred interactions of the predator-prey system shows that all methods (correlation, MI, CCM and OIF) can work for inferring causality between two variables with different level of granularity. However, considering our definition of predictive causality, as non-linear predictability of diverse events from others predictors, OIF outperforms all other models due to the explicit consideration of synchronization, divergence and diversity of events that define model sensitivity, uncertainty and complexity.

To analyze OIF performance for "low complexity" ecosystems we considered the ambiguous dynamics of sardines and anchovies in oceans. On multidecadal time scales, sardines and anchovies present alternating dominance across global fisheries. Although in in appearance a ecological competition seems to exist between these two species (due to the inversely proportional and synchronized abundance changes), the simultaneous fluctuations of sardine and anchovy stocks suggest that they are also influenced by the ocean temperature.

Incompatible hypotheses have been advanced to try to give explanations for this pattern of alternating dominance, unfortunately leaving aside many other species that clearly exist in the ocean and interacting with sardines and anchovies. Some supposed that these two species act in direct and clear competition [68], while others argued that this pattern is just a result of different or opposite fish dynamics in response to common global environmental forces [69]. Results in [70] revealed that in longer time series not only the negative cross-correlation observed in the 20th century disappears, but the correlation with global environmental forces also has been ambiguous. This lack of correlation is however only related to the fact that species are synchronous and environment$\rightarrow$species effects are characterized by relatively small lags. Yet, lack of evident correlation exist but that does exclude causation. Jacobson and MacCall [67] applied two models to this issue and proposed a relationship that SST influences the behavior and population of sardines and anchovies; however, this relationship vanished when applying the analysis to stock assessments from 1992 to 2009. Although all these possible explanations from different points of view are competing, or even unstable, such results can illustrate that causal interactions among sardines, anchovies and SST present features of nonlinear

dynamics. Nonetheless, and more importantly, the conclusion is that both species are weakly interacting and majorly affected by the environment. All interdependencies exist and they just change in terms of normalized magnitude without neglecting the fact that intrinsic interspecies interaction is also modulated by the environment. As shown by the predator-prey mathematical model (Figure 2.3) and real data (Figure 2.5 and 2.7), synchronized species are certainly affected by a third variable (e.g. the environment and other species) that is forcing both in fluctuating at the same time. Heuristically, it is also very unlikely that two species (or variable more generally) are perfectly synchronized unless they are identical. What Figure 2.4 shows is somewhat very affine to the Heisenberg's uncertainty principle that marks a clear break from the classical deterministic view of the universe. We cannot know the present state of the world in full detail (such as for the "red" dynamics), let alone predict the future with absolute precision. Determinism driven by synchrony allows us to know the current state of the system if unaltered but no future states. Vice versa, uncertainty driven asynchronicity and divergence allow us to predict likely future more than actual present and that appears to be in contradiction to deterministic views but not to realistic probabilistic (or relativistic) view of system dynamics. For this sardines-anchovies "problem" unfortunately the whole complexity of ecosystems has never been considered despite other species may have a dominant effect on their abundance. This underlines as well the importance of scale (biological or otherwise) in framing the problem and bounding conclusions of model results: any "causation" is in reality an interdependence between species bounded by the chosen scales.

In addition to testing OIF on the simplified sardine-anchovy ecosystem, we apply OIF and CCM to a multispecies ecosystem in which 14 dominant fish and 1 jellyfish species were monitored in an abundance census in the Maizuru Bay, Japan. In this ecosystem, all species can be interconnected, leading to an intricate causality system that is extremely hard to estimate considering intrinsic biological species interactions, interactions related to environmental influence, and biological interaction mediated by the environment. "True causality" assessment is also extremely hard because there is no knowledge of which ecological or biological marker can capture all these interaction types. However, when causality is shifted to predictability of patterns of interest, the issue of inferring causality becomes practical and meaningful. Predictive causality between species X and Y, for instance, depends on whether X can assist in predicting the future of Y beyond the extent to which Y itself predicts its own future, and complementarily whether the model can predict

the collective behavior of the system which can be reflected by macroecological indicators dependent on all predictive causality. In this case study, both CCM and OIF models are effective for causality detection from different points of view and they majorly differ considering interaction gradient and computational complexity. As for the latter OIF and CCM have 3 and 20 parameters to populate and the speed of inference assessment for the Maizuru 15 species ecosystem is 2 and 15 minutes, respectively. An already published work used the same time-series data to study how to infer the network and forecast the system stability for the fish community using CCM [21]. Ushio et al. (2018) [21] used a "S-map" model after before using CCM to reconstruct abundance data by stripping the hypothesized seasonality effect modeled as a sine function simplistically. Here we believe that any environmental forcing is important to be captured and affect non-linearly and in an unpredictable way the interactions among species; pure biological interactions are utopianly impossible to measure and they are always context dependent. While it is true that synchronization driven by seasonality can lead to misidentification of "biological" or "true" causality (false negative without the consideration of time lags, or false positive as in [21] is lags are considered in the phase-space) as stated in [21], we believe that the environment is precisely the identifiable cause of synchronization of species in a predictive causality purview. Additionally this is also a more realistic analysis of ecosystems where the environment is central in shaping interconnected populations and then community patterns via complex non-linear function vs. simple sinusoidal seasonality. Lastly, in our opinion another pitfall of the S-map of [21] is even the fact that seasonality importance is weighted for each species in isolation whereas seasonality is also affected interactions in ecosystems; interactions that are suppose to be inferred from data as they are since data already contained non-linear affect of environment and other species dependencies plus single species adaptation.

OIF, through the inference of a better gradient of systemic interaction "causality", predicts how biodiversity changes over time with average value, fluctuations and trend that is closer to the taxonomic α-diversity. This is for effective α diversity with the optimal threshold on interactions maximizing the similarity with observed α (via maximization of Mutual Information). The concept of effective α is very useful because it allows to see which set of interactions is determining levels of α diversity that is potentially more or less sensitive to environmental forcing. For example, S6 shows that high interaction species form a small portion of community diversity that is increasing over time vs. the systemic decrease in diversity (observed in Fig, 8). More importantly, the increasing fluctuations of estimated α from

OIF show the potential way in which climate and/or other anthropogenic changes negatively affects biodiversity in the region considered in relation to intensified interspecies interactions as suggested also in other studies [71–73]. These results are certainly beneficial for fishery resources management and habitat protection aiming to preservation of the fish community with ecological, economic and social outcomes. Thus, models like OIF should be evaluated in this bigger perspective of ecosystem utility or ecological engineering with multiple utilities rather than just seeing these models for the hard inference of pure "biological" interaction causality. Supplementary Information contains further discussions about CCM, TE causality inference, predictability, ecosystem organization and stability.

2.5 Conclusions

Causality detection is a fundamental step in the inference of complex networks with the aim of understanding processes of observed complex systems. This is incredibly important for poorly observable large scale ecosystems whose structural and functional networks are their backbone. However, quantifying the "truly causal" interactions in complex systems is illusory and perhaps impossible to achieve due to data and model limitations (e.g. sampling over space and time), partial ignorance about underlying processes, the strong unmeasurable influence of environmental dynamics, and more importantly their relativity dependent on the scale of analysis and the patterns for which interactions are relevant for. Nonetheless, when causality is shifted to predictability, this issue becomes practical and useful because it links causal *predictable* interactions to some patterns to predict. Patterns that are defining the socio-ecological outcomes of interest for which interactions are signatures of the underlying processes. In this paper we propose the Optimal Information Flow (OIF) model and assess its validity and performance in causality inference by comparing OIF inference to well-documented CCM and correlation model. This is done for a deterministic predator-prey mathematical model, a data-driven sardine-anchovy species dynamics, and an observed multiple fish species ecosystem. We show that OIF, like CCM, is able to effectively identify asymmetric causal interactions between any pair of species. Moreover, OIF performs better than CCM because it provides: (i) a larger gradient of interaction values, yet defining interactions at higher resolution with better definition of asymmetrical interdependencies; (ii) smaller fluctuations around the estimated interaction values for any time delay u,

yet a less uncertain inference; (iii) the estimated memory of one and pairs of species in terms of time delay (without considering future modifications of CCM [74]); (iv) independence on the length of historical data and no requirement for convergence, as well as lower computational complexity (leading to lower sensitivity and uncertainty in state estimates); and, (v) more accurate predictions of temporal changes in macroecological indicators of ecosystems such as for the effective α-diversity after optimal MI-based threshold selection. However, OIF requires the identification of the optimal u value as shown in [31] but this is easily automated by exploring the delay that maximizes MI. Even though a time delay can be defined for any pair of species, we show the average time delay, derived from analyzing all species pairs, can be a global optimum providing accurate macroecological predictions for the ecosystem considered. Thus, the assumption-free information-theoretic OIF is a strong candidate model for the inference of predictable causality in complex ecosystems. A model that is itself an ecosystem mimicking the information flow constituting the backbone of real ecosystems of any nature, from environmental to socio-technological systems. The complexity of real world systems might be higher than the ones studied in this paper, considering the velocity of transitions in rapidly changing systems. Nonetheless, we believe that the dynamics encompassed in our study reflects the fundamental stochastic processes observable in the real world, particularly at stationarity but changes in network topology can be mapped by inferring dynamical networks over time. In a broader uncertainty propagation perspective interactions should be considered as "cooperation" and "competition" between species (or variables more generally) meaning that they contribute in a similar or opposite way to the uncertainty (or information) propagation. Competition means that one variable is predominant (or very diverse) over the other in terms of magnitude of interactions since TE is proportional to the magnitude rather than the frequency of interactions. Interactions that are specifically proportional to the divergence and asynchrony of variables/species which leads to higher predictability. In conclusion, our model can find useful applications in research and applied work for ecosystems at multiple biological scales. A myriad of other models have been proposed in literature, and these can be used simultaneously in real-life applications, to provide the full range of possible states of interactions and average systems' patterns trajectories. As causality is considered as non-linear predictability of diverse events of populations or communities, we believe OIF is the optimal model able to predict the largest divergence of trajectories due to the full consideration of ecosystem states via species probability distribution functions. Predictive

causality is a convenient definition for any ecosystem, or data science problem more generally. However, for investigations of causality aiming to learn underlying physical processes of observed patterns, or for solving pressing issues of real complex ecosystems, a more in depth inquiry of complexity and dynamics (in relation to the target objectives), system learning and stakeholder collaboration are of paramount importance since data and models alone cannot reveal the full picture nor identify realistic and optimal solutions.

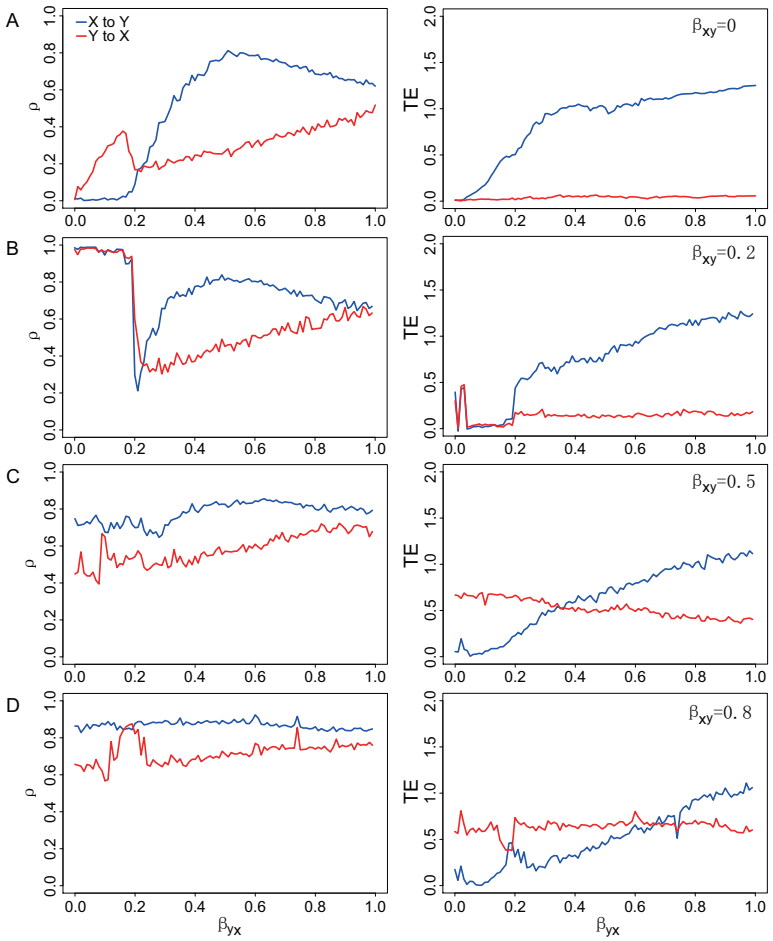

Figure 2.2: **Inferred predictable causality via CCM and TE for embedded true causality.** CCM correlation coefficient (ρ, left plots) and Transfer Entropy (TE, right plots) are shown for the bio-inspired mathematical model in Eq. 2.1 representing bidirectional interactions. The mathematical model indicated as S(β_{xy},β_{yx}) is simplified as a univariate function because β_{xy} is fixed while β_{yx} is free and varying within the range $[0, 1]$. β_{xy} and β_{yx} are establishing true causality while ρ and TE are indicators of predictable causality. Y's causal effects on X is theoretically fixed as a stable value corresponding to each β_{xy}. The greater β_{xy} the stronger Y affects X (estimated by ρ_{yx} and TE_{yx} in red lines). (A) $\beta_{xy} = 0$ means that Y does not affect X and then X dynamics is only related to stochastic dynamics due to birth-death process as in the model (Eq. 2.1). X's effects on Y depends on the value of β_{yx}, theoretically leading to increasing functions ρ_{xy} and TE_{xy} (blue lines) when β_{yx} increases; (B) $\beta_{xy} = 0.2$; (C) $\beta_{xy} = 0.5$; and (D) $\beta_{xy} = 0.8$.

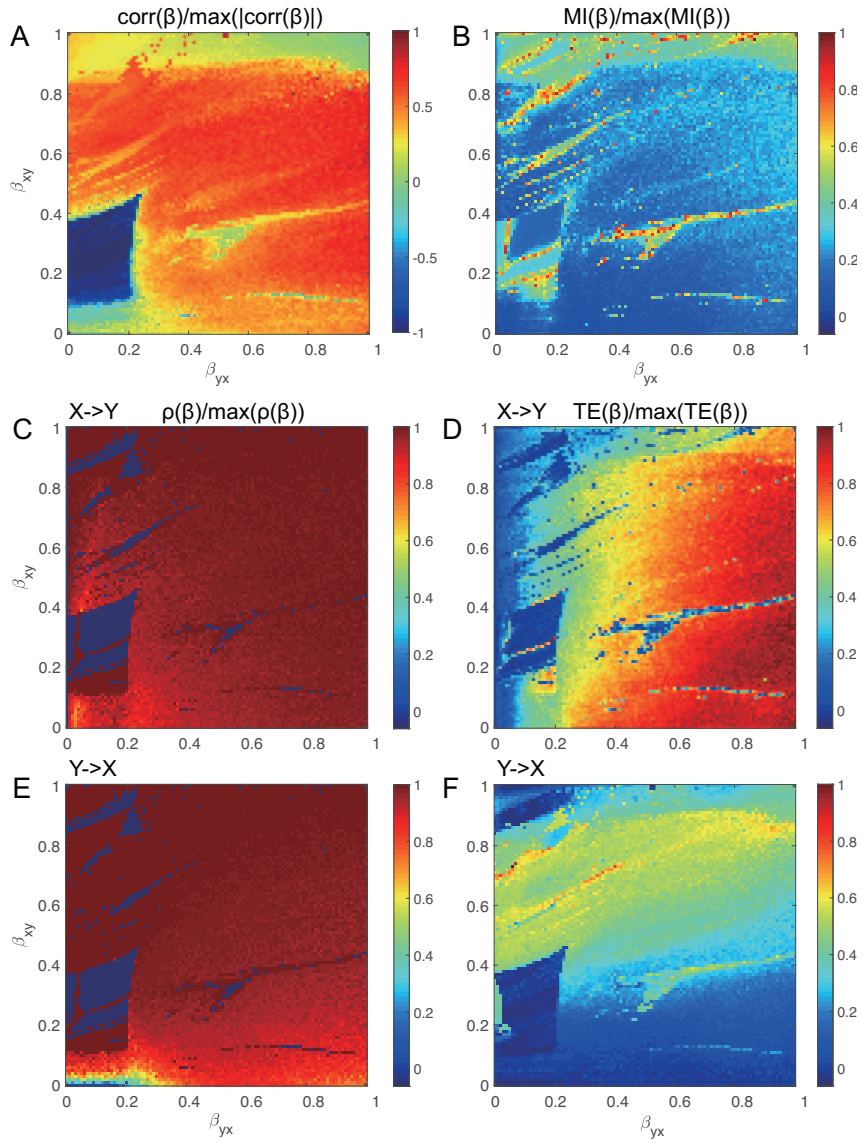

Figure 2.3: **Phase-space maps of normalized coupling predictive causation via correlation, mutual information, CCM and OIF for varying true causal interactions.** Both true causal interactions β_{xy} and β_{yx} are free varying within the range $[0, 1]$, indicating a bivariate model S(β_{xy},β_{yx}) where both species (or variables more generally) are interacting with each other with different strength. (A) normalized correlation coefficient; (B) normalized mutual information; (C) and (E) normalized CCM correlation coefficient (ρ) for interaction directions of $X \to Y$ and $Y \to X$; (D) and (F) normalized transfer entropy (TE) from OIF model for interaction directions of $X \to Y$ and $Y \to X$.

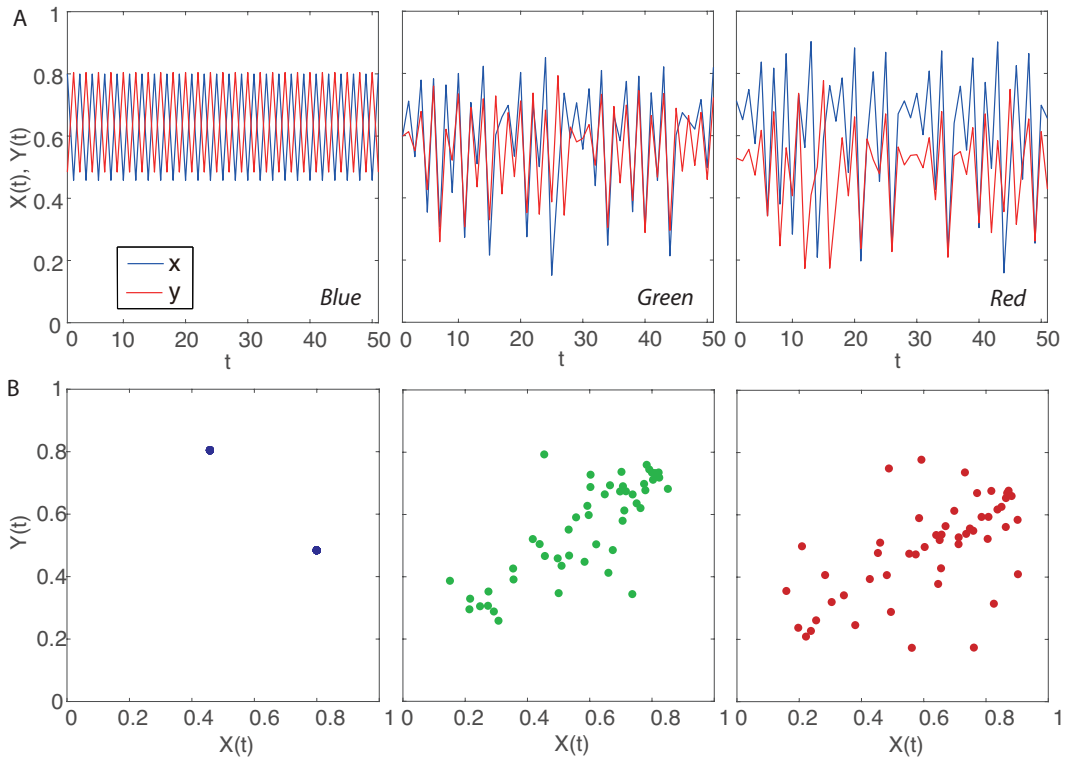

Figure 2.4: **Dynamics of abundance and predictability for the bidirectional two species ecosystem model.** (A) plots refer to the species abundance in time for the mathematical model in Eq. 2.1 for different predictability regimes associated to different interaction dynamics from low to high complexity ecosystem associated to low and high predictability. Blue, green and red refer to a range of predictable interactions as in Figure 2.3: specifically, Blue is for ($\beta_y x$, $\beta_x y$)=(0.18, 0.39) (small mutual interaction, and predominant effect of Y on X), Green is for (0.64, 0.57) (high mutual interactions, and slightly predominant effect of X on Y), and Red for (0.94, 0.34) (high mutual interactions, and predominant effect of X on Y). (B) phase-space plots showing the non-time delayed associations between X and Y corresponding to synchronous and homogeneous, mildly asynchronous and divergent, and asynchronous and divergent dynamics. The transition from synchronous/small interactions to asynchronous/high interaction lead to a transition from modular to nested ecosystem interactions when more than one species exist (Figure 2.6).

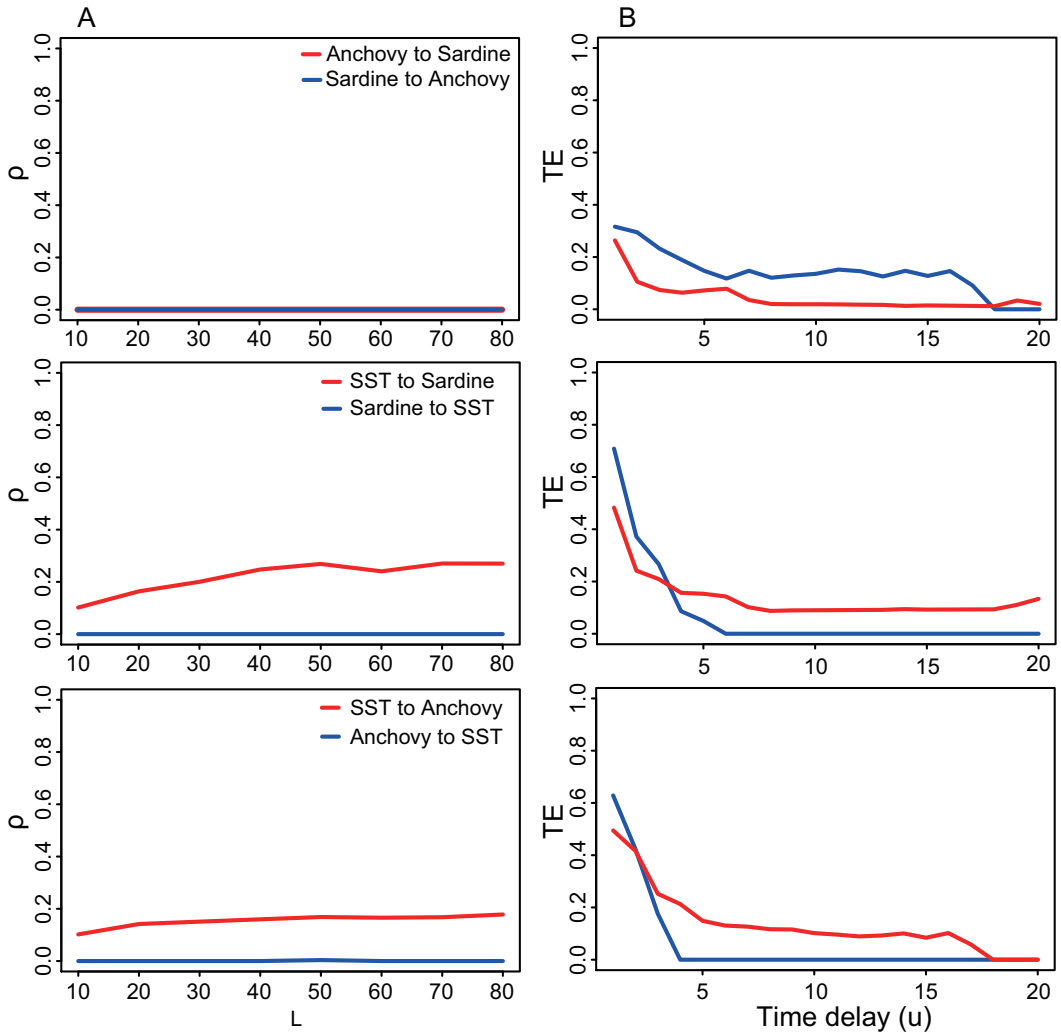

Figure 2.5: **Inferred predictive causality for the sardine-anchovy-Sea Surface Temperature ecosystem.** CCM correlation coefficient (ρ) and OIF predictor (TE) are shown in the left and middle plots for different pairs considered (sardine-anchovy, sardine and SST, anchovy and SST from top to bottom).

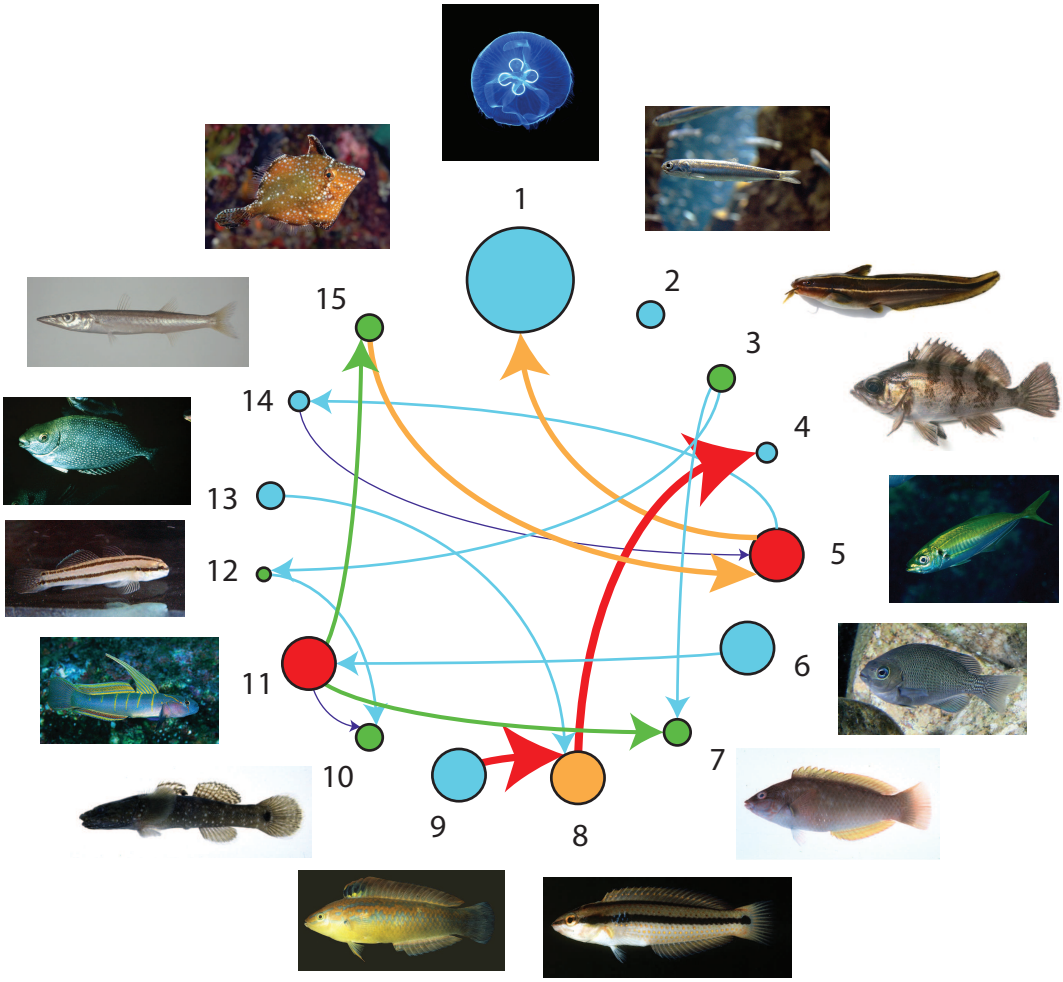

Figure 2.6: **Part of the estimated species interaction network for the Maizuru Bay ecosystem.** Species properties are reported in Table A.1. The color and width of links is proportional to the magnitude of TE (Table A.2); for the former a red-blue scale is adopted where the red/blue is for the highest/lowest TEs. The diameter is proportional to the Shannon entropy of the species abundance (Table A.3) that is directly proportional to the degree of uniformity of the abundance pdf and the diversity of abundance values (e.g., the higher the zero abundance instances the lower the entropy). The color of nodes is proportional to the structural node degree, i.e. how many species are interconnected to others after considering only the CCM derived largest interactions (see Figure 2.7). Other interactions exist between species as reported in Figure 2.7. TE is on average proportional to ρ (Figure A.4 and A.5).

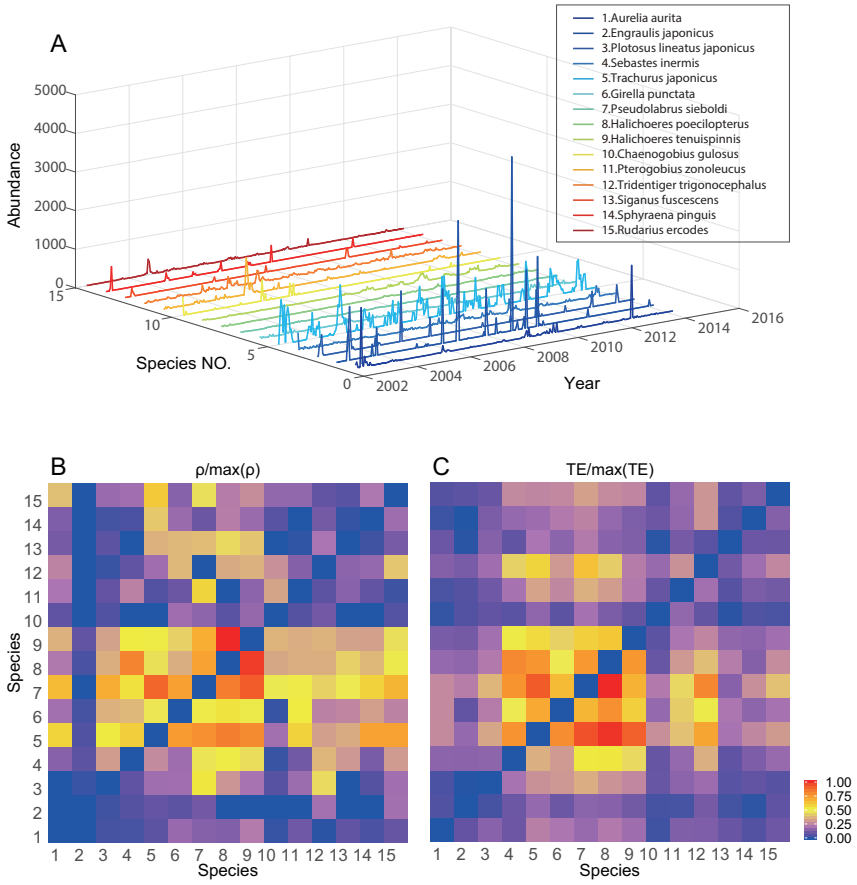

Figure 2.7: **Normalized species interactions matrices inferred by CCM and OIF models for Maizuru Bay ecosystem.** In the census of the aquatic community, 15 fish species were counted in total. Interaction inferential models use time lagged abundance magnitude (CCM) or pdfs of abundance (OIF) shown in A. (B) normalized CCM correlation coefficients (ρ) between all possible pairs of species. (C) normalized transfer entropies (TEs) between all pairs of species from the OIF model. Both CCM and OIF predict that the most interacting species (in terms of magnitude rather than frequency) are 7, 8 and 9 on average. Thus, interaction matrices are more proportional to the asynchronicity than the divergence of species in terms of abundance pdf, although abundance value range defines the uncertainty (and diversity) for each species that ultimately affects entropy and interactions (e.g., if one species have many zero abundance instances or many equivalent values, such as species 2, TEs of that species are expected to be low due to lower uncertainty despite the asynchrony and divergence).

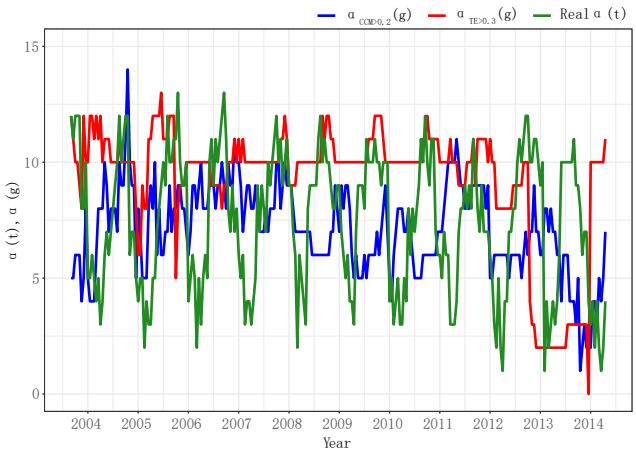

Figure 2.8: **Predicted α-diversity via optimal interaction threshold for CCM's ρ and OIF's TE versus taxonomic diversity.** Effective α diversity from CCM and OIF are shown (blue and red) for an optimal threshold of ρ and TE (i.e., 0.2 and 0.3) that maximizes the correlation coefficient and Mutual Information (MI) between α_{CCM} or α_{TE} and the taxonomic α, respectively. The maximization of the correlation coefficient and MI guarantees that the estimated effective α are the closest to the taxonomic α diversity.

Chapter 3

Optimal Microbiome Networks: Macroecology and Criticality

3.1 Introduction

3.1.1 Microbiome Dynamics and Health

Microbial ecology has become an important topic for health sciences and other basic and applied sciences such as biology, ecology, forensics and agriculture. In particular, the microbiome seems particularly important for ecosystem health in a broader sense, being the primary connector among multiple species, ecosystem structure, functions and services [75]. Recent work has shown how each person maintains a fairly unique microbial fingerprint, and that microbial dysbioses are often associated with shifts in health-status. These shifts are typically associated with the gut that is the most diverse part of the human body considering the bacteria holobiont [76, 77]. We recognize that our microbiota is highly dynamic, and that this dynamics is linked to environmental and individual states [77]. The field of microbiome science is still in its infancy and it is not yet settled upon whether gut microbial community structure varies continuously or if it jumps between "discrete" community states, and whether these states are in common across individuals. In particular, some researchers suggest that gut communities can be binned into discrete enterotypes [78], while others argue that gut communities vary along multidimensional continua without any universality [79]. If the ultimate goal of microbiome research is to improve human health by engineering the ecology of the gut, and other applications are also of interest, we must first understand how and

why our microbiota varies in time and space, whether these dynamics are consistent across humans, whether we can define stable or healthy dynamics, and how these states are associated to the environment. This line of research is primarily missing how microbial diversity is organized considering all its facets and how this diversity changes when species interaction networks change. For instance, the same level of diversity can be achieved via different network topologies that may lead to different health states [80].

3.1.2 Microbiome Diversity and Functional Network Organization

To determine the network organization of the microbiome and associate that to healthy or unhealthy states, we consider Irritable Bowel Syndrome (IBS) as the template syndrome to characterize microbiome dynamics [81]. IBS shows common symptoms of cramping, abdominal pain and diarrhea related to altered gut flora. Previous research has found that the microbiome in people with IBS differs from that in healthy people [81]; however, nobody has demonstrated how the microbiome network is different for these healthy and unhealthy individual groups (i.e., "states" generally speaking when not focused on a particular subpopulation) and how the transition from one to another occurs. By exploring this topic, we propose novel network inferential models for gathering microbiome networks from species big data; these models are based on the principle of maximum entropy that tries to gather the most informative set of variables about stable state patterns with the least amount (but most diverse set) of information [30, 82]. An example can be about sets of species abundance for predicting a diverse set of potential species interaction networks. "Big data" is not only related to the size of the data used but also to the number of calculations required to infer the underlying networks. These computations increase exponentially with the number of species/nodes n considered beyond the geometrical criteria, where the number of connections is $n(n-1)$ in the case of an undirected topology of the network. A directed topology is for instance found when species interaction networks are non-symmetrical which means that the direct influence of two species does not have the same magnitude for different directions of interaction [83]. A variety of different models have been proposed to infer network structures from small and large datasets. For biological systems in particular, the inference of causal interactions among systems' components is a daunting task because not all interactions are known, nor the "true" magnitude

of interactions, considering the data used to assess these interactions and the models [84, 85]. For instance, microbiome networks are in principle different if the used input data are species occurrence, relative species abundance (RSA), geographic range or other features. In addition, for this motivation, we employed assumption free inference models that consider the whole probability distribution of species dynamics and these models were validated considering their ability to predict population biodiversity patterns over time. We extracted optimal microbiome networks as optimal information networks (OINs) [30] for healthy, transitory and unhealthy groups to investigate general patterns and drivers underlying microbiome stability and the interactions among different species in terms of network topology, magnitude and preferential direction. Additionally, we characterized macroecological functions α-, β- and γ-diversity, which describe the temporal organization of microbiome biodiversity considering point time, intertemporal and total diversity. We show how these functions are related to microbiome network features and different topologies emerge for different diversity/health states. The linkage between microbiome networks and macroecology (in particular information theoretic and biodiversity functions) is unique and offers additional insights into the ecology and the evolution of the microbiome with relevance to ecosystem health.

3.1.3 Microbiome Inference, Neutrality and Criticality

Speculations about the underlying processes of ecosystems' organization have been made in the past considering diversity patterns and models able to predict these patterns such as neutral models [86–89], niche models [90–92], and other models such as Lotka–Volterra models based on non-linear ordinary differential equations [93]. Neutral models posit that biological diversity is driven solely by ecological drift without a strong interference of environmental biases that lead to preferential dynamics ("niche") for some species versus others. Neutral patterns exhibit species indicators (e.g., RSA) of all sizes simultaneously without a preferential size. From neutral to niche states, a critical transition is typically observed where species network organization exhibits scale-free behavior [90, 94–97]. This scale-free behavior was thought to occur only at the critical transition point but recent evidence shows that criticality (defined by the scale invariance of ecosystem function reflected by a Pareto distribution) [98] also exists for stable states where system's component organization is optimal due to optimal information sharing among components and the environment [89, 99]. Transitions in network functions are also observed for

neural systems where subcritical and supercritical regimes are defined as the ones corresponding to weakly connected random networks and hyperconnected scale-free networks [100, 101] that can associate to pathologies. These transitions were previously found for geophysical networks and coupled ecological networks [102] for instance, where energy dissipation tends to a global minimum.

Some indications that microorganism cooccurrence patterns are shaped by species interactions that are altered from niche to neutral is available [90]. This also has conceptual and numerical confirmation when thinking and simulating species that are just responding to local resources and species that are somehow "equal" and responding to fundamental speciation-dispersal processes. The former are interacting more randomly with limited dispersal ranges while the latter are interacting with much larger dispersal ranges. The corresponding probability distributions of species diversity for the former and latter cases are exponential and power-law, respectively, corresponding to random and scale-free species networks. Without introducing any model (but with the knowledge of the underlying potential macroprocesses) these changes in network topologies have been observed for large scale ecosystems and other single population systems where topologies correspond to system's pathologies.

However, these models of microbiome characterization are typically driven by some "hard" assumptions about the species interaction network, which may lead to erroneous conclusions about the predicted patterns: in other words, predictability (under some assumptions) of biological patterns does not imply causality considering the hypothesized and implemented processes [20]. Leaving aside the causality investigation, models of microbiome network inference exist [103] and [93]) but they simply infer species co-occurrence networks without assessing the magnitude and directionality of potential species interdependence. A different approach is achieved by pattern-oriented models charactering systems' dynamics [104–106] such as the one here proposed, which do not assume any preferential mechanism a priori but consider the whole information content in data (via probability distributions and their relevance to predict patterns via entropic functions [82]) to claim underlying processes. In this sense we move our discussion of the problem of understanding microbiome dynamics toward one that identifies which information is critical, and how that model criticality [82] is associated to biological criticality [98] also considering the neutrality of biodiversity dynamics. Therefore, rather than trying to untangle biological complexity via fitting some biologically inspired models, we use all data available to check their information content to define all possible

microbiome states and associated diversity patterns. In this information theoretic framework, in particular we show how criticality coincides with neutrality and optimal microbial network organization that leads to healthy states. We also show how criticality corresponds to a scale-free functional networks relating RSA interdependencies even when the functional co-occurrence network of species is not scale-free (this place some warning about inferring networks just based on occurrence data).

As a caveat, it should be noted that neutral patterns does not necessarily imply neutral processes [107] despite many papers try to define one from the other [90, 91]. Furthermore neutral models can predict non-neutral processes (therefore care must be placed when considering predictability vs. causality) and neutrality might not be present at all scales of biological organization [91]. The focus here is on microbiome pattern detection and its predictability, which we believe to be extremely important and the starting point for a top-down investigation of the underlying processes and causality. Different patterns are evident for different health states when RSA interdependence networks are considered, and these networks seem to shape microbiome diversity in many ways considering local, intertemporal and total diversity.

3.2 Methods

3.2.1 Microbiome Data

We considered microbiome data originally published by [108] and later used by [81] for which species data of six individuals are available over time (30 days). Fine scale species Operational Taxonomic Unit (OTU) RSA data were derived by published 16S rRNA and shotgun metagenomic sequencing (SMS) data pertaining to the gut microbiotas. In [108], species-level phylotypes were defined at 97% of sequence identity, which is the lowest taxonomic rank used to identify differences in biological states of interest (e.g., healthy and unhealthy). Two individuals suffered from IBS, two were healthy, one was treated with antibiotics and one was on the verge of being unhealthy. Thus, these two individuals are representative of a transitory state with different directions, from unhealthy to healthy and from healthy to unhealthy, respectively. Durbn et al. (2013) [108] considered the healthy subjects as those individuals who did not suffer from lab-confirmed IBS, and took the patients who had this disease as individuals with perturbations from the healthy state without a priori categorization. In the dataset [108], the healthy period is from time

points before the IBS triggering event altering the microbiome. More specifically, the datasets are composed by two healthy individuals (Individuals A and B in the original datasets [81, 108]), two transitory individuals (C and C1), and two patients with IBS (P1 and P2). The length of RSA data for these individuals are 30 days for A, 15 days for B, 15 days for C, 9 days for C1, 9 days P1, and 14 days for P2.

3.2.2 Time Series Reconstruction

The raw data available present the challenge of individuals whose species abundance is sampled for different time lengths. Computationally, to have datasets with the same length and merge them into one group, we used the method of Least Common Multiple (LCM) for time series reconstruction. LCM extends time series at their maximum feasible length by preserving their probability distribution functions (pdfs); in our case, the pdfs are associated to each RSA and are the inputs for the network inference model that requires time series with the same length [59]. The extended length is the smallest number that is a multiple of the length of original time series of each individual. In this way, LCM guarantees to have the largest dataset representative of the stochastic dynamics analyzed. We calculated LCM considering the number of data for each individual health group. This implies extending the time series at the length of LCM or to maintain the data length if the length of the raw data is equal to LCM. In our study, LCM between Individuals A and B was 30; thus, the length of the abundance time series for A was unchanged while B became 30 (B was repeated twice). This was done by copying the data in B until the 30th day. LCM for C and C1 was 45; thus, both C and C1 time series were extended to 45. LCM for P1 and P2 was 126; thus, both time series were expanded to the 126th day. These examples show that data rich sample are preserved as they are while data poor samples are extended. To create pdfs of RSA representative of each group, we considered the average values of RSA for common species. If for individuals belonging to the same group different species were found, the pdf of RSA was based on the time series as they were. This choice was dictated by the desire to emphasize common dynamics for each group when possible.

3.2.3 Probabilistic Characterization of the Microbiome

We characterized probabilistically the distribution of microbiome macroecological and species interaction network variables (generally indicated as Y as for a generic

random variable) considering the following general exceedance probability distribution function [109]:

$$P(Y \geq y) \sim \begin{cases} \mathrm{e}^{-\lambda_1 \, y} & \text{for } y < Y^* \\ y^{-\epsilon+1} \; f\left(\frac{y}{m}\right) \; \mathrm{e}^{-\lambda_2 \, y} & \text{for } y \geq Y^* \end{cases}, \tag{3.1}$$

where Y^* is the truncation point ("hard truncation") for which the transition in the regime of the probability distribution is observed from exponential to power-law. We refer to "hard truncation" when the pdf clearly exhibits two regimes (for $y < Y^*$ and $y > Y^*$) in which two diverse pdfs can be identified. λ factors are scale factors for the exponential distribution (related to random networks), either above or below the lower/upper cutoff defining the scale-free regime with power-law distribution (associated to scale free networks). m is the upper cutoff after which finite size effects occur faster than exponential decays. We introduce the function $f(y/m)$ to give more generality to the cutoff (or homogeneity) function [109]. $y^{-\epsilon+1}$ is the scaling function where ϵ is the scaling exponent of the power-law distribution; this exponent is a critical exponent associated to the fractal dimension of the process analyzed, yet it is representative of the process dynamics [109]. Note that the probability distribution function $p(y) \; y^{-\epsilon}$ scales with ϵ only. ϵ dictates how the mean and the variance behave, in fact it is related to the Taylor's law scaling exponent [81]. For $\epsilon = 2$, the pdf is the classical Zipf's law that is found for many socio-ecological systems [109, 110].

3.2.4 Network Inference and Dynamical Species Characterization

Information Balance and Exchange

To infer species interaction networks based on microbial RSA data, we based our approach on the model developed in [30] as well as on previous computational efforts [59, 111]. We considered the microbiome as a dynamic network of species interactions (sensu RSA interdependence vs. true causality) where the total free energy and corresponding entropy change over time. The pdf of each RSA for each group was derived by putting together the RSA time series for all individuals; in this network, the RSA was treated as a random variable meaningful of the group and each individual was offering one realization of the same random variable. The RSA matrix was created with compositions in mind and therefore the sum of each sample

was constrained. Considering information entropy as the total dissipated energy's counterpart, the total network entropy can be written as:

$$H(N) \approx \sum_i H(x_i) + \sum_i \sum_{j \neq i} TE_i(x_i, x_j) + \sigma(N) \tag{3.2}$$

where x_i denotes the $i - s$ variables that contribute to the total information of the network N. In our case, x is the RSA of species. In this equation, $H(x_i)$ denotes Shannon entropy, and $TE(x_i, x_j)$ denotes Transfer Entropy from the first variable to the second variable [30, 111]; in our case, both variables are the RSA of two different species. Equation (3.2) represents a fundamental principle of information balance independently of the chosen entropy analytics and forms the general basis of sensitivity analyses. Equation (3.2) states that the total network entropy can be decomposed into the entropy of each individual node plus the entropy of interactions. The sum of absolute TEs is a proxy of the Mutual Information (MI) of a variable, thus it considers the whole set of variable interdependencies; in Equation (3.2), we consider the sign of TE because H(N) should consider the typology of interactions with their sign. $\sigma(N)$ is a noise term that captures the unexplained variability of N related to variables not considered and other discretization factors related to the numerical methods employed in solving the model. Shannon entropy is representative of the species information content (attached to the pdf of RSA) for the whole network and it allows comparing all species in a common framework. Equation (3.2) can also be extended in space if spatially explicit calculations are needed, as in [30]. Note that H(N) is inversely proportional to the free energy of the system so the lower H(N) the higher the free energy and the higher the total dissipated energy. Evolution self-organizes systems toward states where H(N) is minimized.

The computation of TE was based on the distributions of the two variables of interest (i.e., RSA) conditioned on their histories. Comparing the conditional probability of the variable on its own history with the conditional probability of the variable on both its own history and the history of a predictor variable provides asymmetry in determining predictive abilities of one variable onto another. Thus, a directed network can be inferred. According to equation 2.4, directed TE of two time series variables, denoted as X_i and X_j, was calculated as $TE_{X_i \rightarrow X_j}(\tau)$, where $X_{i,\tau}$ and $X_{j,\tau}$ denote the respective histories of X_i and X_j at time t as well as considering all past values for the period $t - \tau$. Here, we consider the same memory lag for X_i and X_j but in principle historical dependencies can be different when

considering other variables and the variable itself. In our microbiome study, X_i and X_j are RSA of species i and j. This definition is the most general definition of TE and neither conflates dyadic and polyadic relationships between species nor assumes any causality.

The definition of TE can assume that the processes analyzed obeys a Markov model, which is suitable for memoryless stochastic process. This implies that future states depend only on the current state and not on events that occurred before it. Thus, in a Markov process, it is assumed that $\tau = 1$. This is usually true, especially for rapidly varying processes (such as for microbial RSA); however, this constraint can be relaxed by choosing temporal lags that are small enough to focus on short-term interdependencies which are not related to long dependencies in the underlying processes. In our case, study RSA values of two randomly selected species did not correlate with RSA values for $\tau = 1$; thus, memory processes are relevant and, as in [59], we selected the τ that maximizes the interdependency between two species assessed by the functional distance (see Equation (3.11)). Note that TE, as calculated in Equation (2.4), should be interpreted as information flow vs. information transfer [84] because conditional entropies are used to exclude indirect pairs of species whose interactions is of second order importance. This approach has been criticized by some authors (e.g., [112]) if "causality" is indeed claimed about the inferred interactions and in consideration of the fact that polyadic relationships may be underrepresented. In this study we spouse the view of [112] for which TEs are considered as measures of reduction in uncertainty about one time series given another (thus, with predictable power) with potential but not certain causality, leaving aside the issue of what specific biological causality is investigated (e.g., influence, physical causality, etc.). The idea of using conditional entropies is solely related to find the most informative set of species to identify the core microbiome interaction network.

Maximum Entropy Networks

Subsequently, the inference of interspecies TEs, among all values of TEs the question remains on which value is the most informative about the potential causal relationship between two variables. We emphasize that here "causal" is in the sense of of predictability, sensu uncertainty reduction, rather than "certain" biological reality. As in [30], we proposed to select TEs that lead to the maximum entropy for the inferred network. This corresponds to maximize the Fisher information ma-

trix [113] that produces the lowest complexity and the highest informative set of information about a pattern of interest. MaxEnt favors probability distribution functions with maximum entropy as the most general distributions that fit the observed data [114]. This theory can be applied to a functional network where edge weights are based on TE. The network with the greatest total entropy can be similarly favored as the most general network structure that fits the observed data. The method considers all possible pairs of variables in both directions for predicting a pattern of interest. The edges that comprise the network with the greatest total TE are then included. Selecting the edges that contribute to the greatest amounts of TE, according to the MaxEnt theory, produces the network that most accurately describes "causal" patterns among the included variables. Note that MaxEnt should be interpreted in an information theoretic sense, where higher entropy means higher information. We show how this entropy (useful to characterize the system) is related to the state of each health group that has a more ecological and physical sense in a thermodynamic purview; in particular, how the absolute value of total entropy is lower for stable and healthy states vs. unhealthy ones.

A utility function is needed to establish the function where MaxEnt is applied. The utility function can be thought as a systemic (network) value function $\sum_{i,j} f_{i,j}(X) w_{i,j}$ (potentially multiplied by weight factors $w_{i,j}$) where value functions $f_{i,j}$ are TEs among RSAs. These TEs, as in Equation (2.4), assess the potential causal interactions between species pairs. Thus, the utility function is the total network entropy $H(N)$ (Equation (3.2)) that needs to be optimized in order to define necessary and sufficient TEs with the maximum entropy. The optimization can be subjected to feasibility constraints, for instance related to the ability to control certain species or data limitations. In the context of the present goal of creating a microbiome network indicator, the value functions $f_{i,j}$ are defined as:

$$f_{i,j}(X) = \begin{cases} TE_{X_i \to X_j}, & \text{for } \{X_i, X_j\} \in E_{MENet} \\ 0, & \text{for } \{X_i, X_j\} \notin E_{MENet} \end{cases}, \tag{3.3}$$

where $\{X_i, X_j\}$ represents the directed edge connecting X_i to X_j, and MENet (Maximum Entropy Network) represents the set of directed edges in the network with the maximum total network entropy $H(N)$. The selection of edges to be included in the network is determined by finding the network with the greatest total entropy as in Equation (3.2). In the present study, the utility function was defined as the total TE of the network (plus Shannon entropies of each RSA but those turned out to be second- or third-order factors that can be neglected), and it is maximized

by selection of the $f_{i,j}$ functions. To the best of our knowledge, this is one of the first times that TE was framed in a decision analytical model via a network threshold entropy criteria that defines MENets.

Optimal Information Networks

To reduce redundancy in creating a MENet, variables that are strongly predicted by other variables (hypothetically establishing a strong causality—in a predictive sense rather than in a biological one—if prediction accuracy of one decreases quickly when removing the other [20]) can be excluded. This can be done by evaluating the weighted in-degree and out-degree of each node in the network (i.e., TE). Nodes with a greater weighted out-degree than in-degree can be included in the Optimal Information Network (OIN) that one among many MENets with the same average total entropy. These nodes are strongly predicting the variability of other nodes, thus the overall network dynamics. OIN is then the necessary and sufficient MENet for predicting microbiome function. Here, we refer to microbiome function as the information network related to the interdependence between RSA measured by TE; this function is not the "true" biological function but it is likely related to the variability in mutual abundance that is commonly found in any complex ecological systems [115]. Thus, OINs are purely information networks and not causal biological networks. This entropy reduction to define OINs based on conditional entropies (calculated on sets of potentially influencing species that do not affect much the total entropy, yet removing the indirect interactions as in [111] in order to estimate information flow vs. information transfer [84], where the former is more likely representing "causal" species interactions)) can be further achieved by introducing functions $g(X_i)$, defined as follows

$$g(X_i) = \begin{cases} 1, & \text{for } \sum_j f_{i,j}(X) > \sum_j f_{j,i}(X) \\ 0, & \text{for } \sum_j f_{i,j}(X) \leq \sum_j f_{j,i}(X) \end{cases}, \tag{3.4}$$

where $\sum_j f_{i,j}(X) = OTE$ and $\sum_j f_{j,i}(X) = ITE$. OTE and ITE are the total outgoing and incoming TE for a node, respectively. Thus, variable inclusion depends on the comparison of the TE projected by the variable X_i onto the other variables and the TE projected by the other variables onto X_i.

The defined function g was then used to create the total network entropy that

can be used to carefully describe the network dynamics:

$$H(N \equiv OIN) = \sum_i H(x_i) \cdot g(x_{i,t}) + \sum_i \sum_{j \neq i} TE_i(x_i, x_j) \cdot g(x_{i,t}) + \sigma(Y) \quad (3.5)$$

which represents the sum of all necessary variables that were included by the structure of MENet in a multi-criteria value function, and the sufficient variables after the redundancy exclusion to form OIN. In this way, the OIN inference was based on information theoretic and functional topological criteria to screen: (i) the necessary information to maximize network entropy H(MENet) (i.e., total information content); and (ii) the smallest non-redundant information to sufficiently predict total network function (of maximum entropy H(OIN)). Note that the first criterion on H(MENet) is a global one on the total information content while the criterion on H(OIN)) is a local one on the information of a node with respect to the functionally connected nodes. This entropy minimization is somehow the equivalent of the energy minimization of other optimized networks [116]. However, this OIN is the network with the highest accuracy in predicting macroecological patterns of diversity over time that are dependent on fluctuating RSA. Then, OINs are characterized by the highest information content (lowest uncertainty), highest information diversity (e.g., represented by the values of TEs), and lowest complexity.

Assessment of Species Importance and Collectivity

After the inference of OINs, it is possible to quantify the importance of different species considering their variability in isolation and in cooperation with other species for predicting the dynamics of the microbiome. Species first order importance and interaction for reproducing the network dynamics are then calculated considering new indices based on nodal information flow rather than on Mutual Information Indices (MII). σ_i describes species interaction and is calculated as the ratio between the total Outgoing Transfer Entropy (OTE) as information flow ($OTE(j) = \sum_i TE_{j \to i}$) and the total network entropy, while μ_i describes the species importance as the ratio between the nodal Entropy as information content (using Shannon entropy) and the total network entropy. These Transfer Entropy Indices (TEI) are useful when no systemic variable is needed (contrary to [30]),

and analytically they are formulated as:

$$TEI = \begin{cases} \sigma_i = \dfrac{OTE(j) = \sum_i TE_{j \to i}}{H(OIN)} \\ \mu_i = \dfrac{H(x_i) \cdot g(x_{i,t})}{H(OIN)} \end{cases}, \tag{3.6}$$

When considering a systemic indicator [30], MII are better suited to identify variable importance because no directional influence is needed. MII use the mutual information (MI) normalized by the entropy of the output variable considering one independent variable or pairs of variables for predicting a dependent variable Y that is in this case undefined. These MII indices are $s_i = \frac{MI(X_i;Y)}{H(Y)}$ and $s_{ij} = \frac{MI(X_i;X_j|Y)}{H(Y)}$, where X_i is any variable (e.g., RSA) and Y is the predicted variable built using the same process of constructing OINs but selecting variable features rather than keeping entropy of species as independent variables. The use of TE can give further information about the directionality of causality (in a predictive sense of the model), and the time-lag of the causality.

3.2.5 Macroecological Indicators

To characterize the microbiome as an ecosystem we introduce macroecological indicators that aim to describe ecosystems' collective dynamics of diversity locally, within communities or time points, and globally. In this paper we use such macroecological indicators that are time dependent (because space information is not provided and hardly inferable) and of order zero mathematically speaking (as in [63] the order is related to the exponent to which the probability of RSA is elevated to). For a set of unique distinct species $\mathbf{S} = \{S_1, S_2, ..., S_n\}$ whose RSA $\mathbf{X} = \{X_1, X_2, ..., X_n\}$ changes over time, we define the local species diversity, or α-diversity as:

$$\alpha(t) = \sum_{k=1,t}^{n} p_k(t)^0 \tag{3.7}$$

where $p_k(t)$ is the probability to find one species at time t. Thus, α is the sum of diverse species at any given time during the observation period (30 days) or the reconstructed period (see Section "Time Series Reconstruction"). Considering this definition of α it is easily noticeable that the sum of the entropy of all RSA

$H_\alpha = \sum_k H(x_k) = -\sum_k p_k(t) \log p_k(t)$ is proportional to the Shannon index that is the local species diversity of order one [63].

Leaving aside the controversy about the definition of interspecies diversity over time, i.e., species turnover, we define β-diversity as the complementary variable of species similarity (here introduced via the Jaccard Similarity Index (JSI) as in [102] and [87]):

$$\beta(t) = 1 - JSI(t) = 1 - \frac{S_{t,t+1}}{S_t + S_{t+1} - S_{t,t+1}} \tag{3.8}$$

where $S_{t,t+1} = \sum_{k=1,t}^{n}(p_k(t)^0 + p_k(t+1)^0)/2$ is the number of species present at both time steps if $p_k(t)^0$ and $p_k(t+1)^0$ are $\neq 0$, otherwise $S_{t,t+1} = 1$. $S_t = \sum_{k=1,t}^{n} p_k(t)^0 = \alpha(t)$ is the number of species present at time t (or $t+1$) (Equation (3.7)). Note that, β-diversity as a measure of species turnover overemphasizes the role of rare species as the difference in species composition between two communities or two time steps is likely reflecting the presence and absence of some rare species in the assemblages.

Note that the definition of β in Equation (3.8) is proportional to the "true" β that is classically defined as the number of diverse species between two samples (either over space or time). β-diversity can also be defined as a second order index where the entropy related to β is $H_\beta = H_\gamma - H_\alpha$ [63] where $H_\gamma = H(N)$ is the total network entropy (Equation (3.2)). Considering the variation of diversity over time β-diversity is proportional to the complementary of the mutual information $1 - MI_{X_i,X_j} = 1 - \sum p(X_j, X_i) \cdot \log_2 \left(\frac{p(X_j,X_i)}{p(X_j)\, p(X_i)}\right)$. However, $1 - \beta(t)$ is proportional to the sum of the TEs. These relationships between information theoretic quantities and macroecological indicator is novel and worth being addressed in further papers.

The total diversity γ is defined as:

$$\gamma(t) = \sum_{k=1,t=1}^{S,T} p_k(t)^0 \tag{3.9}$$

that can be established over time or over the total number of speciation events M. M is the sum of all species at any given time independently of their diversity calculated from time $t = 1$ to the final time of observation T; equivalently, M is the number of events when new or existing species are introduced. A speciation event is an event when a species is introduced in the microbiome; this species can be already present or can be a new distinct species that is established over the total number

of speciation events M. The concept of speciation event is introduced because that determines the number of total species introductions independently of the true temporal dimension. Thus, the speciation event focuses on the dynamics of the process independently of time because it counts events. Considering M allows one to map how the total diversity changes as a function of biodiversity meaningful scales, equivalently to the species–area.

vs. mapping its change over time (that may not be an influencing variable). The total number of speciation events can be related to the number of unique species S (i.e., all distinct species occurred in the time period) as follows:

$$M = \sum_{k=1}^{S} m_i x_i^0 \tag{3.10}$$

where S is the number of unique species across the whole observation period, x_i is the RSA of the counted species, and m_i is the number of times that species occurs. Considering the validity of the information balance equation (Equation (3.2)) that leads to the diversity balance equation $H_\gamma = H_\alpha + H_\beta$, the total diversity can also be calculated as $\gamma = \alpha \cdot \beta$ [63].

3.2.6 Functional and Structural Network Metrics

The topological organization of the microbiome is characterized via structural and functional complex network metrics. Functional metrics are based on information theoretic functions that quantify the interactions among species while structural metrics are based on the geometry of the network and can be derived from the former ones.

The functional distance between species is defined as:

$$d_f(X_i, X_j) = \min_\tau \ e^{-MI(X_i(t\pm\tau), X_j(t))} \tag{3.11}$$

where the minimum value of the distance is taken for all possible time delays τ. X_i and X_j are the RSA of species i and j and MI is the mutual information evaluated for different values of the temporal scale of species dependency τ. The τ that minimizes the distance d_f is chosen for capturing the maximum interdependence MI_{max}. Such distance as in [59] quantifies the magnitude of the most meaningful interactions between species in a predictive sense: the higher MI the shorter the distance that signifies high levels of interaction (sensu predictability) without specifying the directionality. Thus, because of the inability of assessing the direction of

interdependence between species (whether that is information transfer or flow [84]), MI (or d_f equivalently) is a metric useful for identifying the most interacting pairs of the microbiome rather than individual species.

The calculation of the structural distance is based on the functional distance and the concept of the shortest path. The structural distance is then defined as the minimum number of steps from one node (species) to another independently of the magnitude of these steps (e.g., in terms of TE). Thus, analytically the structural distance is defined as:

$$d(X_i, X_j) = \text{argmin} \left[\sum_{i,j} d_f(X_i, X_j)^0 \right] \quad \text{if } A_{ij} = 1 \tag{3.12}$$

where $A_{ij} = TE_{ij}^0$ is the adjacency matrix that can be formulated in terms of TE. The rationale for considering the shortest paths is related to the exponentially large ensemble of distances as a function of the number of nodes and the fact that biological systems always optimize information transmission [116]; however, Pareto shortest paths are always chosen [116].

In terms of connectivity, the functional degree is defined for the directed network as the sum of the weighted in- and out-degree (i.e., TE) elevated to a power exponent equal to zero. Then, analytically the functional degree is:

$$k_f = k_{in} + k_{out} = \sum_{i,j} \left[f_{i,j}(X)^0 + f_{j,i}(X)^0 \right] \tag{3.13}$$

where $\sum f_{i,j}(X) = TE_{ij}$ is the transfer entropy as defined in Equation (2.4).

The structural degree is defined by thinking the network as an undirected network (without signs related to TEs), thus

$$k = \sum_i a_{i,j} \tag{3.14}$$

where $a_{i,j} = 1 = TE_{i,j}^0$ if i and j are connected. Classically, the structural degree considers the number of connections independently of the bidirectional pathways implied by TE. Thus, functional degree is always greater or equal to structural degree.

3.3 Results

3.3.1 RSA Analysis

The simplest analysis of the microbiome starts by looking at the temporal trajectories of RSA. By a simple cursory analysis, it was evident that the average RSA of the healthy microbiome is lower than the average RSA of the unhealthy microbiome independently of the species; however, the maximum RSA was found for the healthy microbiome and the species with the highest RSA is one of the most beneficial for health. A recent dataset with absolute abundances suggests that healthy gut microbiota have higher total abundances than diseased ones [117] but no studies exist about the universality of this abundance-health relationship. By looking into species diversity (Figure 3.1A), it was observed that the average number of species at any time point (α) is lower for the healthy microbiome than the unhealthy one. This may seem in contrast with previous findings that report higher diversity for healthy microbiome or in general for healthy ecosystems [95, 118, 119]. A controversy on the subject is already found in literature [118], thus just maximizing total diversity without considering how that diversity grows and is organized is not intuitively a necessary and sufficient ingredient to achieve a stable healthy state. More importantly, the RSA-rank pattern (Figure 3.1B) shows only one dynamical regime, corresponding to the common Zipf–Mandelbrot model for RSA [120], for the healthy microbiome vs. two regimes for the transitory and the unhealthy microbiomes (double Pareto, lognormal or exponential regime). Figure 3.1C shows that the decay in richness over RSA is higher for the unhealthy microbiome; this result underlines the fact that higher diversity does not imply stability because of the suboptimal, yet unsustainable distribution of species in the unhealthy microbiome. Stability is related to network topology [76], which also affects diversity and the systemic fluctuations of the microbiome, as shown by the Taylor's law [81] that highlights how variance in RSA abundance changes with the mean. "Optimal" organization is in this case referring to the healthy state as a reference state because it has the smallest fluctuations for the highest achievable total diversity growth rate γ' (this is the Pareto solution) and the associated network topology is more resilient to random node removal (Figure B.3). The Pareto solution has the largest diversity growth rate and is not by chance accompanied by a Pareto-like species interaction network where interactions are inferred by TE (Figure 3.5b). Figure 3.1B,C shows the RSA-rank plot and the Preston's plot of species diversity dependent on

RSA. The RSA-rank shows two dynamical regimes for the unhealthy and transitory groups: a result that likely confirms the bimodality in local species richness α. By plotting the Preston's plot in log-log, a scaling relationship was found showing a faster decay in species richness for the unhealthy group.

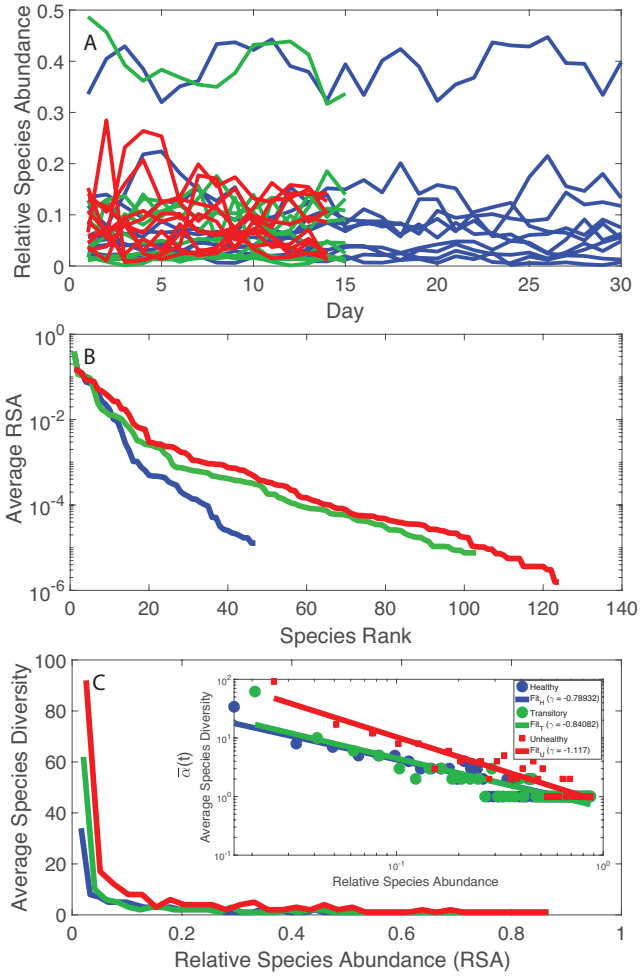

Figure 3.1: **RSA trajectories, RSA-rank, and Relative Species Abundance**. Blue, green and red curves refer to the healthy, transitory and unhealthy microbiome, respectively. The healthy microbiome shows smaller fluctuations in species diversity α vs. RSA and one regime when considering the RSA-rank profile. An inverse scaling law was detected between the average species diversity and RSA (inset (**C**)).

Considering the RSA of species in time, from the most to the least relatively abundant, a transition in the pdf of RSA was observed from a pseudo-normal distribution (corresponding to a homogenous spatial distribution) to a Dirac-like distribu-

tion (corresponding to a singular point distribution) considering the maximum and minimum RSA. Considering the RSA of all species together (Figure B.2) the transition is less dramatic, from an exponential to a log-normal-like distribution. Intermediate RSA species, independently of species belonging to the healthy, unhealthy or transitory group, show a scale-free like distribution underlying the fact that these species are fundamentally important in the function of the complex microbiome as highlighted in [95]. Rare species seem also to display a truncated scale-free behavior (limited by their maximum RSA as a finite size factor rather than limited by spatial biological constraints), which also underlines their importance for the microbiome organization. These pdfs are a signature of species interaction networks for different RSA groups: pseudo-random, scale-free, and small-world topology for the highest, intermediate and lowest RSA class, respectively. Further results discuss the connection between RSA and species information flow.

3.3.2 Network Inference

The inferred microbial networks corresponding to the three microbiome groups are shown in Figure 3.2. Maximum entropy networks evidence the different topology in microbiome organization for healthy, unhealthy and transitory group. In the structure of these networks, the size of each node is proportional to the Shannon entropy of the species and the color is proportional to the structural degree. In Figure B.3, we show the networks whose nodal color is proportional to the total outgoing TE (OTE) that is likely more representative of node activity in a collective network sense. The higher is the value of the structural degree (or OTE in Figure B.3), the warmer is the color. The width of each edge is proportional to the TE between pairs and the direction is corresponding to the directional influence. All OINs are special MaxEnt networks, i.e., networks for which the total network entropy is maximized (MENets) and where redundant nodes are removed (see Section 3.2.4). Thus, OINs allow one to identify the fundamental functional species interactions useful for predicting microbiome dynamics. The transition in network topology, from random to small-world (tending toward a scale-free network) for the unhealthy and healthy groups, is manifested also by the shift in total entropy pattern (left plot in Figure 3.2). The latter is asymmetrical and symmetrical for the random/unhealthy and scale-free/healthy microbiomes, respectively. This type of network transitions has been observed for large ecosystems. The network entropy plots show that network entropy over information flow is roughly symmetrical for healthy individuals, ex-

pressing that the interconnectedness in healthy communities is more dynamically balanced than unhealthy ones. Figure S3 shows microbiome networks for a high value of the threshold on TE_{ij}, which establish the information exchange (of flow) between species above which links become relevant. However, these networks are no more OINs. Considering the total network entropy and its decomposition, it was observed that the most important nodes in terms of OTE (Equation (3.5) and Figure B.6), that is the information flow necessary to predict all other nodes' dynamics, are the dominant species in making up the total information network (Figure B.5). In other words, the entropy of each single node in isolation $H(x_i)$ is a second- or third-order factor in determining the total network entropy.Figure B.7 shows that most species interactions (TEs) are positive for the unhealthy microbiome, which is underlying the evidence that mutualistic positive feedbacks leads to instability; therefore, higher α and γ diversity in short and long term do not guarantee stability if interactions are predominantly in one direction. The healthy microbiome instead has balanced positive and negative interactions that lead to microbiome stability.

Figure 3.3 shows macroecological indicators of diversity of the microbiome for healthy, unhealthy and transitory individuals. We show that species diversity α, and total species diversity γ are the highest in the unhealthy group (for which average RSA is also the highest) but species similarity $1-\beta$ and the diversity growth rate α' over time are the highest for the healthy group. This is a critical result that shapes microbiome organization around healthy or dysbiotic states. The highest fluctuations in RSA and macroecological indicators (in particular, α and γ) were observed for the transitory and unhealthy groups. These results underline the potential conclusion that too high levels of diversity are possibly unsustainable, leading to unhealthy unstable states related to the abnormally excessive multiplication of species in the guy ecosystem. These species may be invasive from outside sources or subspecies created within the gut as a response to external stressors. It is interesting to note that the behavior of the pdf of α informs about the potential states of the microbiome in each group. The pdf is platykurtic multimodal for the unhealthy microbiome, which suggests the presence of multiple unstable states, and it is leptokurtic monomodal for the healthy microbiome which implies one stable state. The transitory microbiome shows an almost symmetrical pdf underlying the fact it exists in between the healthy and unhealthy microbiome. These results underlines the resilience of the microbiome as a whole dictated by the ability to change as a function of external stressors as well as the higher stability of the optimal healthy state. However, the latter seems easy to perturb considering the lower entropy (and

probability, or corresponding high free energy) defined in one state. This ability to change state is also a good indicator of gut adaptability and human body resilience.

Species collective interaction and singular importance is shown in Figure 3.4 by plotting the information theoretic TEI σ_i and μ_i (see Methods, Section "Assessment of Species Importance and Collectivity"). The top 10 interacting species are also the least relatively abundant for the healthy microbiome and the most detrimental; however, these species are controlled by other species and the microbiome is organized into a healthy state. Figure B.7 shows that from the top to the least 10 TE species there is a shift in the pdf of RSA from a bimodal to a monomodal distribution for the healthy microbiome. For the transitory and unhealthy microbiome, instead, there is a shift from a leptokurtic (Dirac-like) to a platykurtic pdf (uniform-like). The top 10 TE species are the most detrimental bacteria ("antibiotic") but their RSA is small for the healthy microbiome; this means that these bacteria are controlled (in terms of RSA variability) by all other beneficial bacteria. The top 10 TE species are mostly characterized by positive interactions (positive TEs) while the least ten 10 TE species are characterized by negative interactions (feedbacks). For characterizing species collectivity or single species dynamics, as well as for predictability, OTE that is a node function is better suited than TE that is a link function. The pdfs of OTE in Figure B.6 show more clearly the changes in species dynamics for each health state and overall species activity manifested by the magnitude of OTE. The top 10 OTE species are always characterized by positive feedbacks vs. the least 10 OTE species with negative feedbacks (top and bottom plots of Figure B.6). Figure B.8, by plotting the pdf of all TEs and OTEs for any group, further emphasizes the fact that there is a positive bias and an asymmetry for the unhealthy group species interactions.

The non-linear duality between microbiome structure and function is shown in Figure 3.5 where structure is considered via the network degree (Figures B.9 and B.10) and function is about the nodal information flow OTE. The epdfs show how microbiome function is much more suited to show functional network topology versus microbiome structure. Function is a much more important property than structure which is just based on geometrical analyses of cooccurrence species networks. This scale-free function may be related to the scale-free behavior of the intermediate RSA species, as shown in Figure B.2. As shown in Figure 3.2, visually, the healthy microbiome functional network is tending toward a scale-free topological organization. Statistics of the functional scale-free network based on TE are in Figure 3.5. This mild scale-free organization (see, e.g., [121], where the authors highlighted

the difficulty in defining the classification for these networks into one topology radically) does not correspond to a scale-free distribution of α-diversity (Figure 3.5 bottom plot) that instead is exponential. Additionally, some functional network features beyond the inferred RSA-based interdependence (TE and OTE) show a bimodal or Poisson distribution (Figure B.10) characterizing more small-world networks rather than scale-free ones. However, we point out how these features are more structural than functional (see Equations (3.11) and (3.13)) since they characterize species interactions directly. The non-linearity among structure, function and microbiome service (i.e., diversity in this paper) is highlighted when plotting α dependent on functional network degree and distance (Figure B.10). α diversity increases for high values of the functional degree (Equation (3.13)) but does not have a clear trend when considering the functional distance (Equation (3.11)). $\alpha(d_f)$ is lower for the unhealthy than the healthy microbiome for the same range of functional distances which highlights the more random distribution of diversity in any dysbiotic state. We observed 72, 378, and 9647 unique values of functional distance for the healthy, transitory and unhealthy group. The highest diversity in functional distances for the unhealthy group confirm the fact that the unhealthy microbiome is more densely connected and the number of small distances (high species interdependencies) is lower than the healthy one. However, the healthy microbiome is more clusterized into species clusters. The values of functional distance were normalized and the distribution of α over the normalized distance shows a random arrangement for the unhealthy group with respect to the healthy one (Figure B.10).

We found the most interesting results when we combined microbiome service and function indicators, for instance considering total macroecological diversity γ and OTE. Figure 3.6 shows the relationship between γ and the temporal sampling scale (i.e., the number of speciation events) in analogy to the species–area relationship widely used in macroecology. The plot shows a scaling relationship valid for two orders of magnitude whose exponent is higher for the healthy than unhealthy group underlying the optimal growth of diversity for the healthy microbiome. Considering this optimal diversity growth relationship, it is meaningful how the transitory microbiome has the largest value of γ' leading to a change in diversity from the healthy species "poor" to the unhealthy species "rich" microbiome. These results are in synchrony with the power-law decay of species similarity $1 - \beta$ over time (Figure 3.6, bottom left). When considering OTE of species as a function of their RSA, we found a surprising scaling law over four orders of magnitude; this law with an average exponent close to 1/4 (very common in biology, for instance the

mass-specific Kleiber's law [122]) implies a decay in species interaction for highly relatively abundant species. When comparing γ over OTE (Figure 3.6D), a non-linear growth is detected where a common increase in total diversity occurs until a critical species interaction value, above which γ slows down or remains stationary, at least for the healthy and transitory groups. For the unhealthy group, the growth of γ seems to slow down but not reach a stationary state; this may relate to the continuous multiplicative generation of detrimental species in the gut.

3.4 Discussion

We employed an information theoretic model for the inference of microbial species interaction networks based on RSA interdependence. The model was used to infer microbial networks associated to different health states and is suitable for predicting selected biodiversity patterns characterizing the space-time organization of bacteria α-, β-, and γ-diversity. Thus, the primary purpose of the model is not to infer causal (or "true") species–species interactions among bacteria. The computational inference of "real" interactions is always very hard—provided that there is a complete knowledge of the reality on which results can be validated—and any inferred interaction is always dependent on the analytics and data used. For instance, RSA profile may not necessarily contain the information about all species–species interactions aimed to be assessed but still the question remains about what is truly an interaction (aimed to be measured) since any physical or functional interaction may not necessarily reflect any change in RSA, or other biomarker. Additionally, any change in RSA or other biomarkers may be related to other external factors, such as environmental fluctuations, which alter species simultaneously. What is certainly true, however, is that, if the inference model detects strikingly different patterns for different population groups, then those patterns likely tell something meaningful about different dynamics and collective environmentally driven changes [104, 105].

In this perspective the entropy-based model is focused on the predictability of patterns vs. causal investigation of mechanisms. The proposed model can be applied to both abundance and RSA, or other biomarkers, without any special modification. Theoretically, the pdf of abundance and relative abundance is the same leaving aside numerical artifacts; independently of this, RSA seems better suited for this type of ecological analyses because it informs about changes of species abundance with respect to the whole community. Abundance and/or RSA seems also the most

likely to detect species functional roles and interactions as highlighted by recent studies [115]. Constructing a network for each health group is the purpose of studies such as ours that try to identify common group dynamics in populations independently of individual variability, where universal group dynamics in microbiome is the core quest). The identified network topologies have a correspondence with the dynamics of RSA, that is a critical dynamics for the scale-free information network associated to the healthy state, and exponential dynamics for the random network associated to the unhealthy state. The total network entropy is the lowest for the healthy microbiome for any threshold of the information flow TE (Figure 3.2). This implies higher free energy available to the healthy microbiome and lower information needed to function where information entropy in the physical space can be thought of as the average interspecies communication/interdependence. The lower entropy in species collective interactions has certain implications for data collection, potentially implying fewer data are needed for characterizing healthy microbiomes. This is because one single globally stable state was identified for the healthy microbiome (in the entropy pattern in Figure 3.2) vs. multiple stable states for the unhealthy microbiome (one globally and two locally stable state for high, medium and low value of network entropy, respectively). These states correspond to different biodiversity states in terms of α, β and γ. The existence of multiple dysbiotic states seems to confirm the previously observed "Anna Karenina effect" [119] where "all healthy microbiome look alike, instead each unhealthy microbiome is diverse in its own way". More theoretically speaking, the lowest entropy across the system's landscape of potential states is a sign of criticality that is the state toward which any ecosystem tends to [99]; the critical state is where there is a balance of system's self-organization and environmental influence [105].

The inferred patterns in this paper are representative of confirmed health states where individuals are confirmed representative samples ([108] published the original dataset) for IBS and non-IBS people, as reported by [81]. Patterns and methods are proposed to highlight what is relevant to look at when describing state transitions and characterizing health states. The number of individuals sampled in a population matters as a function of expected or reported patterns' changes. Reliability is not only dependent on the sample size but also on the consistency and differences within and among samples. In this particular study, we found striking differences between potential health states and many times concordant with the reported literature. Further research is required to test the biological universality or local specificity of these patters across a much larger population sample than the

one considered. Analyses were made considering varying data lengths for individuals, which did not change any pattern considered significantly. This means that the dynamics represented in the time series is well contained at least in the smallest data sample available. The smallest reliable sample is for ten data points that seems in this case the minimum data length to have in order to have representative probability distributions.

Considering the issue of compositionality, which is related to the issue of having samples consisting of proportions of various species with a sum constrained to a constant, the theory suggests that a small number of species should increase compositional effects. In our case, the number of species is 47 at minimum and that should limit the effect of compositionality because the sample is large enough. Microbiome sequence datasets are typically high dimensional, with the number of species much greater than the number of samples. The consideration of pdfs limits the issue of compositionality, as well as the focus on group vs. individual statistics limits the issue of data sparsity (considering both rare species and the length of time series). Of course, this macroecological purview does not imply any strict causality inference but rather aims to set up the basis for the predictability of microbiome group features. This is also because there is no well established data or model to identify what is truly a causal effect between species, although some advancements have been made in the field of information theory such as in [84] where information flow (such as the one used in our model via TE after entropy reduction) proves to assess local causality vs. information transfer via simple TE. Arguments have also been formulated about the general validity of TE to infer causality (see [112]). However, beyond these analytics centered debates, the fundamental argument should also be focusing on what kind of interaction based on data is truly inferred, what is the interaction that is wished to be inferred, and what is the modeler choice of analytics selected to represent reality [85]. All these elements of discussion would make the interpretation of results clearer, such as the distinction between inferred networks for predicting patterns vs. inferred networks claimed to represent the physics of the biological system considered. Despite sophisticated approaches to statistical transformation (such as centered log-ratio transformation that can remove the constraint of the sum of species proportions), the analysis of compositional data may remain a partially intractable problem because RSA is the information that is available. Given these findings, promising work has been done on addressing compositional data as a significant challenge to co-occurrence network inference, but the problem is still not solved. However, TE is not affected by compositional data (provided

enough data are given to characterize pdfs) precisely because it uses pdfs in network inference and the pdf of RSA, raw abundance, and any transformation applied to all species is the same. A problem may arise only when data are asymmetrically transformed in a way that the pdf of one or more species is altered.

The entropy/free energy patterns (or "entropy-flow patterns") in Figure 3.2 do not show any strong scale invariance as for instance in [30], likely because no pure scale-free networks are observed in the microbiome organization. In this study, we focused on the total entropy as a utility function versus the value function defined in [30] (based on a systemic indicator) where raw values of network variables were considered rather than TEs among them. The focus on network variable interdependence (that is between species in this context) rather than nodal values (i.e., RSA for the microbiome) leads to a higher variability in network entropy patterns. Therefore, we believe that the focus should be on network function in order to better characterize networks; this is substantiated by the higher importance of species interactions (OTE) versus species independent dynamics (represented by nodal entropy), as shown inFigure B.5. This figure shows that OTE makes up almost the whole Network Entropy (H_N) (see Equation (3.5)) so Nodal Entropy has little importance. Entropy-flow patterns are then useful for detecting scale-invariance in the functional topology of the network and for identifying MaxEnt states. Additionally the entropy-flow patterns can reveal healthy vs. unhealthy states by considering the symmetry of the entropy distribution; if symmetrical positive and negative species interactions (TEs) are found these interactions sum up to zero leading to a healthy neutral state. The asymmetry of unhealthy microbiome can certainly relate to non-neutral states created by strong stressors, as highlighted theoretically in [113]; these state may not allow host individuals to keep the microbiome "on a leash" [123] that causes overgrowth of abundance and multiplication of species. However, the broken symmetry can be indeed manifesting an unhealthy state. The neutral state also coincides with the critical state because of the tendency of the network toward a scale-free organization manifested by the epdf of OTE (Figure 3.5), higher functional distances and smaller functional degrees (Figure B.10).

To assess the robustness of microbiome networks, we considered the network topology for high thresholds values of the interspecies TE. In other words, we considered as meaningful TEs, only those above a certain threshold. According to the 80-20 Pareto principle (that states that 20% of subcomponents make up at least 80% of a system's dynamics [124]) (note that this principle works for scale-free systems), we considered only the highest 20% of TEs for the inferred networks. These

Pareto high threshold networks show that the healthy group maintains the topology while changing TE; this is because healthy networks are more scale-free than unhealthy ones (see Figure 3.5, middle, for the epdf of OTE), yet scale-invariance is preserved when changing the threshold defining the scale at which the network is constructed (or observed). This scale analysis is equivalent to make experiments when random nodes are removed simulating a random attack on networks [125]; thus, we can also claim the higher resilience of the healthy network for the microbiome. However, this result is expected considering the known optimality of scale-free networks [116]. The scale-free configuration enhances stability as confirmed by the calculation of the dominant eigenvalue for both the adjacency and TE matrices; the dominant eigenvalue is the smallest for the healthy group that is a signature of network stability [76].

The "non-pure" scale-free organization of the microbiome confers the ability to adapt to different externally-driven changes and to adapt vs. a more stable scale-free topology. Overall, we suggest to focus on TE and OTE as the best indicators of microbiome function (for pairs and node functional characterization), vs. any other indicator, since those are related to species interdependence. As highlighted in recent studies (see [115]) abundance determines the functional role of bacterial phylotypes in complex communities; rare and common bacteria are implicated in fundamentally different types of ecosystem function [115]. Such knowledge could be used, for example, to understand how bacteria modulate biogeochemical cycles, and to engineer bacterial communities to optimize desirable functional processes. Microbiome service is here identified by any microbiome diversity indicator in analogy to how services are also expressed for large scale ecosystems. Certainly, it is true that α-, β- and γ-diversity cannot be "equated" to large scale ecosystem services (i.e., the benefits that people derive from nature and how these are quantified as "natural capital"), but any diversity measure is a valuable indicator of biological function at any scale of biological organization (see, for instance, [126] and [127]) much more than structural indicators, as shown in this paper. Therefore, there is a desired ecosystem service-function nexus that is desirable and related to healthy states (which is the benefit individuals get from having the "right" value and patterns of macroecological indicators manifesting optimal biodiversity organization). Of course, especially in microbial ecology where the identification of species is more difficult than large scale ecosystems, there are arguments about the utility and validity of different diversity metrics such as γ vs. evenness. Nonetheless, independently of this, we argue that our analyses would result in equivalent conclusions.

For instance, in our case, high γ corresponds to low evenness and vice versa; thus, biodiversity patterns would reveal opposite trends but provide the same meaning because of the γ-evenness relationship.

In our microbiome data, we considered the complementary of β-diversity over time via the Jaccard Similarity Index (JSI) and we showed that JSI is higher for the healthy than the unhealthy microbiome over time. This means that the local species richness, α, tends to be more equal to previous values over time; however, this underlines the stability of α (species organization) in the healthy state. For the unhealthy microbiome, the similarity over time is lower (i.e., higher species turnover, or higher β-diversity) such as for the corals in [119] that are evaluated over time as a function of external stressors. In other types of ecosystems, e.g., in coral ecosystems under stress, [119] found that the true β-diversity increases over time. In macroecology, leaving aside the debates about the many definitions of species turnover, and in an entropic context the true β-diversity is the ratio between regional (γ) and local species diversity (α) [63]. This definition is in line with the general information balance equation (Equation (3.2)) and the more specific diversity balance equation $H_\gamma = H_\alpha + H_\beta$ as in [63]. An increase in β is typically associated with a decrease in α as much as we observe for the healthy microbiome, and this is also associated to fluctuations of α that are smaller than those for the unhealthy microbiome. The "proportional species turnover" (i.e., where $\beta_p = 1 - \alpha/\gamma$, when considering γ partitioned into additive rather than multiplicative components) that quantifies what proportion of species diversity is not contained in an average representative sample, is also higher. This emphasizes how our results are robust independently of the peculiar definition of species diversity indicators. In ecology these quantities are typically evaluated over space and in healthy conditions 1-β has a relatively fast decay but never goes to zero; this means that heterogeneity exists but even communities far apart have species in common. Considering space in unhealthy conditions, typically the "true" β-diversity is smaller than in healthy conditions because much more homogeneity is achieved. However, heterogeneity is a good thing as shown for ecosystems at any scale of biological organization.

The higher variability of β-diversity in healthy individuals highlights the "Anna Karenina phenomenon" for human microbiomes. The principles underlying the phenomenon states that dysbiotic individuals vary more in microbial community composition than healthy individuals paralleling Leo Tolstoy's dictum that all happy families look alike ("each unhappy family is unhappy in its own way"). The stability-unimodal pattern of diversity is concordant with current theories looking into β-

diversity vs. solely α-diversity for the stability of ecosystems [118]. This is also concordant with the network entropy pattern that is unimodally stable for the healthy group. Thus, we innovatively highlight the linkage between information exchange and diversity in biological systems. [128] previously found that ecosystem hotspots are those that maximize the Value of Information (of biodiversity) which coincides with those that minimize β-diversity variability over time. The multiplicity of "unhappy/unhealthy" states is reflected by the network topology that is random for the unhealthy group, which allows many more potential unhealthy microbiome combinations. We support the position of previous studies that Anna Karenina effects are a common and important response of animal microbiomes to stressors that reduce the ability of the host or its microbiome to regulate community composition. These effects may be transient and necessary to bring back the system to the healthy state.

Similar to other ecosystems, we show that scale-invariance (that is occurring for the healthy microbiome) does not arise from an underlying criticality (where fluctuations becomes bigger and bigger causing the system to tip abruptly) nor self-organization at the edge of a phase transition. Instead, it emerges from the fact that perturbations to the system exhibit a neutral drift (also relate to small extrinsic environmental changes) with respect to the endogenous spontaneous dynamics. This *neutral* dynamics, similar to the one in genetics and ecology, shows fluctuations of all sizes simultaneously that likely determine power-law distributed species diversity (as well as power-law information exchange among species). The tipping point that was observed, i.e., between healthy and unhealthy microbiome, is a second-order critical transition where exogenous fluctuations are too large to be assimilated by the system and the microbiome tips from healthy to unhealthy. This transition is evident in the shape of the pdf of microbiome function and diversity (as microbiome service) but not in the shape of microbiome structure (unless a rescaling in size is performed, for instance for the microbial network degree; see Figure 3.5).

The introduction of new pathogens driven by the environment can lead to the alteration of the whole ecosystem microbiome [81]. In our case study, despite the non-explicit consideration of the disturbance agent, we found a transition in IBS individuals from healthy to unhealthy states. However, this disturbance agent was considered by [108] and [81], who worked on the original dataset. Independently of the disturbance, healthy individuals have larger gradients of speciation events and higher growth rate for γ-diversity because they produce more species (diverse of not) to guarantee necessary/basic biological function and other functions related to extreme fluctuations. Not all species need to be present all the time and that is

likely the motivation for which the average γ' is higher for healthy and transitory individuals than unhealthy people as well the average γ is lower for healthy ones. γ' seems to reflect the general dynamical systems' pattern indicated by the Heap's law [129] that regulates the rate of diversity produced by a system. This is associated to the Taylor's law regulating mean and fluctuations and the Zipf's law (in our case of RSA which influence macroecological indicators). In a more ecological purview, the species–area-like relationship in Figure 3.6A can also emphasize the island biogeographic effect where for islands/healthy individuals γ is lower but γ' is higher than the mainland/unhealthy people due to optimal growth (ideally not impacted by invasions). The higher γ for unhealthy individuals is likely related to invasive species for instance attributable to external sources; healthy individuals instead, have a gut flora composed by only endemic species. In a general view, Taylor's law regulating RSA fluctuations, Zipf's law governing RSA distribution, Heap's law relating γ's growth over time, and the mass-specific Kleiber's law are all liked together by the Pareto optimal principle of self-organized design [130–133] that can inform about the optimality or pathology of biological systems.

The microbiome in the gut is similar to any ecosystem: no other species at all scales of biological organization can survive optimally if the microbiome is altered. The microbiome is the linkage between the fundamental genetic organization of life and the stochastic environmental dynamics; in the context of a person's growth, it is possible to refer to those two processes as nature and nurture. The proposed information theoretic global sensitivity and uncertainty analyses (Figure 3.4) allow one to map the dynamics of species considering their interactions and absolute influence, and to see how these quantities vary considering their intrinsic biological variability and environmentally driven variability. One must keep in mind that these interactions are based on mutual RSA interdependence assessed by TE, so TEs might not represent the whole "true" interactions among species; however, recent evidence points to this conclusion [115] but there is still a lot work to be done in this area. In the healthy state, more species (fewer in number) are influencing the collective dynamics with a more organized distribution of interactions ("hierarchical" organization), while for the transitory and unhealthy state all species (higher in number) are somehow behaving equally and likely driven by external environmental stimuli ("random" organization). This organization is also reflected by network properties (Figures B.9 and B.10) that can be altered for the same set of species/diversity. Researchers have found that cooperation promotes ecosystem biodiversity, which in turn increases its stability without any fine tuning of species

interaction strengths or of the self-interactions (i.e., neutrality) [134]. Even small values of TEs (close to zero) manifesting mutualistic interactions (positive) among species can stabilize the dynamics. Stability increases with the ecosystem simplicity where the latter is related to the scale-free like organization of bacteria. On the other side, too much cooperation (e.g., dictated by networks for high values of TE) promotes instability and complex random networks. It is interesting to note that this scale-free cooperation of species leads to Taylor's laws [135] between mean and variance of RSA where Taylor's exponent is different for healthy and unhealthy groups [81]. However, this reemphasizes the connection between time dynamics, network organization, and ecological patterns of diversity and RSA [97, 134, 136]. In particular, it has been shown that higher-order interactions (e.g., captured by σ_i in our model) have a stabilizing role [136]. These higher-order interactions are all those beyond the simple pairwise interactions whose sum indeed cannot explain the whole composition and dynamics of ecosystems [137]. We show that these higher-order interactions cannot be prevalent because some species must have an independent dynamics (captured by μ_i) otherwise instability and tendency toward disorganized unhealthy state is very likely (Figure 3.4).The healthy critical state is in fact characterized by an heterogenous distribution of σ_i and μ_i for species that is optimal for the microbiome.

The definitions of detrimental and beneficial bacteria (some of them listed in Figure 3.4) were based on previously published papers. For instance, *Lactobacillaceae* and *AcidobacteriaGp18* are beneficial, while *Neisseriaceae* and *Campylobacter aceae* are detrimental. Of course, this is just a rough categorical classification because as we emphasize in this work, for a bacteria being detrimental or not is a function of relative abundance and network topology rather than just being present or not in the microbiome or other independent properties without considering the bacteria collectivity. Microbiome functional network topology defines how all bacteria behave synergistically and that synergy brings a healthy or an unhealthy state. Additionally, the functional topology characterization, for instance determined by OTE, can avoid the issue of determining precisely what true "species" are that is a debated topic in microbial ecology. The focus is on portfolios of interacting species whose interaction is responsible for the microbiome dynamics/state. This result sheds some light into a vision where a diminishing role of network hubs (considering total information flow) is reported as found by other studies [138]. The least relatively abundant species for the unhealthy microbiome are the most interactive and the least detrimental. On the contrary, the most relatively abundant species

(Figure B.4) for the unhealthy microbiome are the least interactive and the most detrimental. These analyses considering the activity of species show the importance of weak ties (interactions) for the healthy and unhealthy groups. This is in accordance to general dynamical principles such as the Granovetter principle about the strength of weak ties for the systemic dynamics of a complex system [139]. For the healthy microbiome, the highest RSA species interact the least and these species are the most beneficial. These species–specific analyses, when verified, are useful for detecting species that are more beneficial or detrimental and this knowledge can lead to design probiotic treatment, microbiome transplants [140], and large scale ecosystem microbiome controls [141] for instance.

Universality in human microbiota dynamics, whether present, can be ideally manipulated in a similar or even identical fashion in multiple individuals for population health. Following the discovery of universality and the demonstration of beneficiary effects of specific interventions, microbiome engineering efforts can be applied to a large number of people. In this way, microbiome engineering will be highly cost-effective as a public-health based approach. This is in sharp contrast to the excessive cost of "precision-medicine" approaches that try to target individual microbiome dynamics by considering it as a purely individual-based feature. Current frontier topics are also related to the understanding of how the microbiome and functional brain networks "communicate" [142]. It seems that the nervous system contribute to dictate which microbes inhabit the gut; this in turns affects emotional response and long term well being beyond short-term health. The hypothalamic–pituitary–adrenal axis (HPA axis) is a primary mechanism by which the brain can communicate with the gut to help control digestion through the action of hormones [142]. It seems that the nervous system, through its ability to affect gut transit time and mucus secretion, can help dictate which microbes inhabit the gut, which in turns affects emotional response and long-term well being beyond short-term health.

3.5 Conclusion

An information theoretic model for the inference of microbiome networks and the related biodiversity organization over time is proposed. The model consists in the assessment of transfer entropy-based species interactions after entropy reduction calculations that remove the second-order indirect interactions between species as in the works of [84] and [111]. Maximum entropy networks are then extracted

considering the highest information content without model overfit; overfitting is avoided by removing the redundant variables for the simplest MENet, that is an Optimal Information Network. Species interactions should be interpreted in terms of species predictability rather than causal mechanisms due to the data- and model based-dependence of the inferred interactions [112]. The macroecological validation of the model was performed considering the ability to simultaneously predict the pdf of α-diversity, γ-diversity growth, species similarity $(1-\beta)$ decay, and the RSA-rank profile. This validation allowed predicting other biodiversity patterns such as the Preston's plot of average species richness dependent on species RSA. Considering the application of the model to healthy and IBS symptomatic individuals, the following points are worth mentioning without lack of generality.

- Directed species interdependencies and phase transitions of the microbiome over time were detected. The healthy microbiome is characterized by balanced positive and negative species interactions vs. the unhealthy microbiome where most species interactions are positive. The balanced interactions were evidenced by the symmetrical pattern of the total network entropy as a function of the pairwise information flow (TE) vs. the positively biased asymmetrical pattern of the dysbiotic microbiome. The healthy symmetrical network entropy pattern underlines the neutral "sum to zero" dynamics of species interactions (based on RSA); the same neutrality was found for biodiversity of large scale ecosystems at stationarity that are driven predominantly by intrinsic ecological stochasticity (ecological drift). On the contrary, unhealthy microbiome entropic patterns are affected by environmental disturbances; the positive bias in information flow (that may relate to infections and antibiotics, as shown in the original data [108]) causes an overgrowth in RSA of many opportunistic species as well as the generation of new detrimental species. The categorization of beneficial and detrimental species was based on published literature; however, we emphasize how important it is to consider collective bacteria topology vs. individual bacteria behavior when defining health and disease;

- The healthy state is characterized by the highest total species diversity growth rate γ' (leaving aside the transitory microbiome) and the lowest loss of species similarity over time, i.e., species turnover ($(1-\beta)'$). A relationship similar to the species–area relationship for large scale ecosystems was found between γ-diversity and the number of species generations with an exponent equal to

0.20 on average. The fact that the healthy microbiome has the lowest average total diversity (γ) is in contrast to what is observed in large-scale ecosystems at stationarity where the highest total diversity correspond to the stable and supposedly healthy state [143]. However, we speculate that an optimal diversity growth is oriented toward maximizing growth rate rather than total diversity (as according to many Pareto portfolio theories). The latter can lead to over-redundancy of microbial interactions and instability as observed for the dysbiotic microbiome; the highest γ diversity for unhealthy ecosystems is related to non-endemic species. Hence, we tend to challenge the diversity–health–stability hypothesis when for diversity the total systemic diversity γ is solely considered without the consideration of "invasive" species and γ';

- We observed a phase transition of the second order from the healthy to the unhealthy state and vice versa. The transition from healthy to unhealthy is characterized by typical signs of transitions observed in many complex systems [144], i.e., an increase and a decrease in mean and variance of species diversity while approaching the transition ("critical slowing down"). In the unhealthy state the variance of α is higher than in the healthy state and concentrated around two values which underline the likely chaotic-like dynamics of the microbiome. In terms of microbiome functional network topology, a transition between the scale-free to the random network topology is observed. The critical state, defined by a scale-free-like organization of microbial species interactions, coincides with the neutral state (i.e., for the symmetrical network entropy pattern) emphasizing how criticality does not necessarily occur at critical phase transitions, particularly for second-order transitions as in this case. Rather, criticality can coincide with neutrality in open energy dissipative systems, as observed in other complex systems [89]. Criticality at the phase transition can favor gut adaptability but may pose high risks to tip to unhealthy states. Neutrality implies lower topological complexity and higher dynamical stability (corresponding to higher symmetry, higher organized information exchange, lower entropy/total information, higher diversity, and higher predictability (or information content)) considering the scale-free and small-world functional and structural organization of the microbial network. We emphasize how the healthy local stable state is dynamically flexible because of the lower entropy (i.e., higher free energy) and more predictable due to the more organized collective behavior of species; however,

due to the gradient in entropy moving from locally stable unhealthy conditions to the globally healthy stable one is hard;

- A probabilistic linkage was found between microbiome function and services, defined by species interaction topology and biodiversity organization, respectively. We did not find any correspondence between microbiome structure and function, which emphasizes the non-linearity between the two and the importance of assessing function rather than structure in biological networks. We propose the total Outgoing Transfer Entropy (OTE) as the measure to identify the most influential nodes (and pairs); these nodes are able to predict the behavior of all other connected nodes, as well as of the whole microbiome. OTE is largely determining the total entropy of the network compared to the sum of nodal entropies whose contribution is negligible. This emphasizes even more the role of collective behavior vs. individual nodes considered in isolation. The highest OTE nodes have the lowest RSA, and these are the most beneficial and the most detrimental bacteria for the dysbiotic and healthy microbiome. A scaling law was found between OTE and RSA with an exponent close to $1/4$ that is similar to the mass-specific Kleiber's law [122] where the species specific metabolic rate is the OTE and the mass is the RSA. A power-law distribution for the microbiome function (i.e., the sum of nodal OTE) was found for the healthy state (with an exponent $\sim$2 that implies finite mean but infinite variance suggesting how the healthy condition is prone to perturbations enhancing fluctuations of all sizes) despite no information (or resolution) invariance being detected in the network entropy pattern (see [30]). The lack of scale invariance in the entropy/free-energy phase space may imply the metastability of the microbiome that can indicate its resilience in terms of ability to move quickly from one state to another.

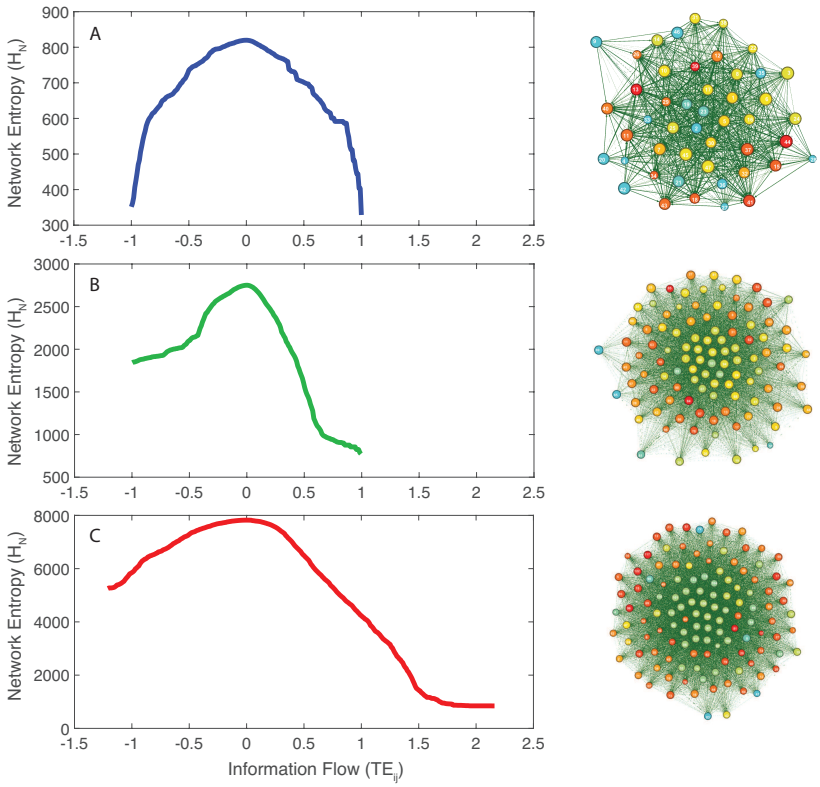

Figure 3.2: **Network entropy patterns and inferred Optimal Microbiome Networks.** Network entropy dependent on the pairwise information flow (TE) (left pattern) and extracted Optimal Information Networks for the microbiome on the right (Maximum Entropy Networks after node redundancy exclusion). The size of each node is proportional to the Shannon Entropy of the species; the color of the node is proportional to the structural degree (in Figure B.3, the color of each node is proportional to the sum of total outgoing TEs of each node (OTE); the higher is the OTE, the warmer is the color); the distance is proportional to $exp(-MI(X,Y))$ where $MI(X,Y)$ is the mutual information between species RSA x and y; the width of each edge is proportional to the pairwise Transfer Entropy; and the direction is related to $TE(i->j)$; the direction of this edge is from i to j.

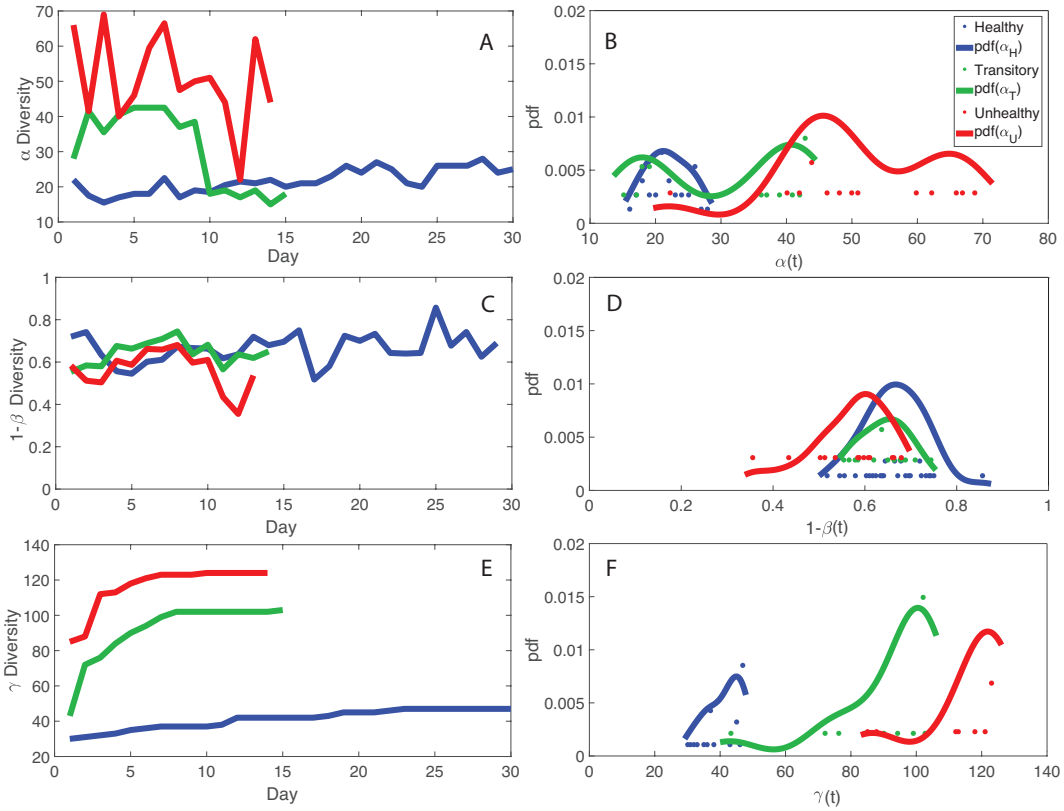

Figure 3.3: **Macroecological indicators of microbiome networks and probabilistic characterization.** Average α, species similarity $1-\beta$, and total diversity γ are plotted as a function of time. Their probability distribution is shown on the right.

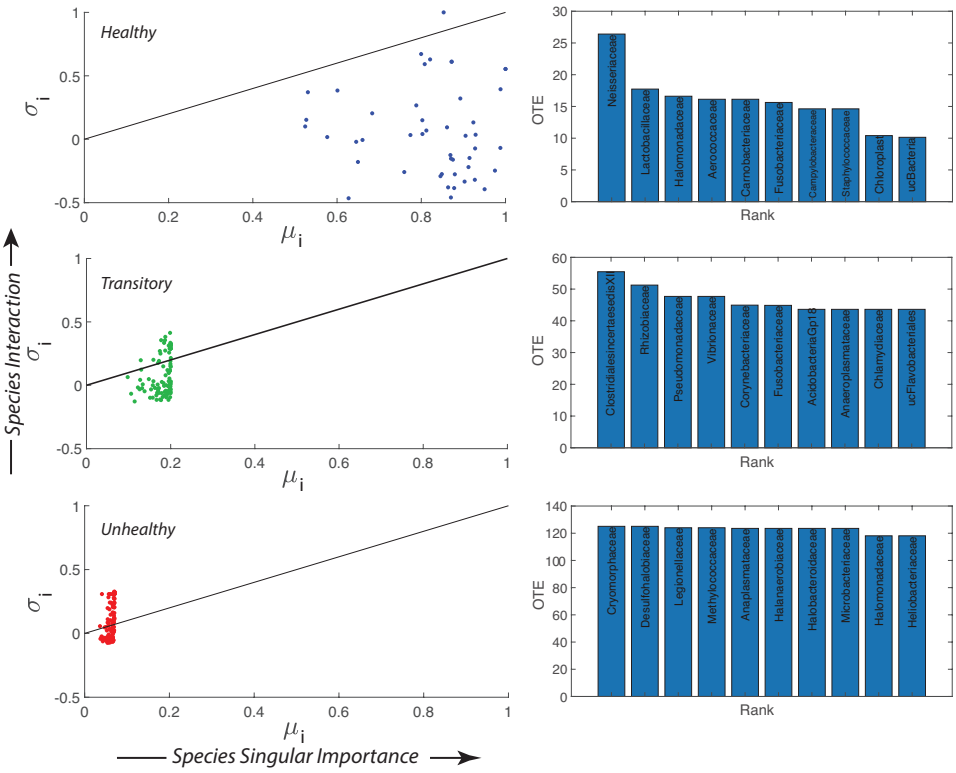

Figure 3.4: **Importance and interaction of microbial species, and top 10 most active species species.** Transfer Entropy Indices: σ is describing species interaction and is calculated as the ratio between the total Outgoing Information Flow (OTE) ($OTE(j) = \sum_i TE_{j \to i}$) and the Total Network Entropy, while μ is describing the species importance as the ratio between the Nodal Entropy (Shannon Entropy) and the Total Network Entropy. The continuous line in each σ-μ plot shows the critical edge that describes a state between regularity and chaos. On the right, the top 10 most active species in terms of OTE (and least relatively abundant) are ranked. These species are the most detrimental for the healthy group and the most beneficial for the unhealthy one.

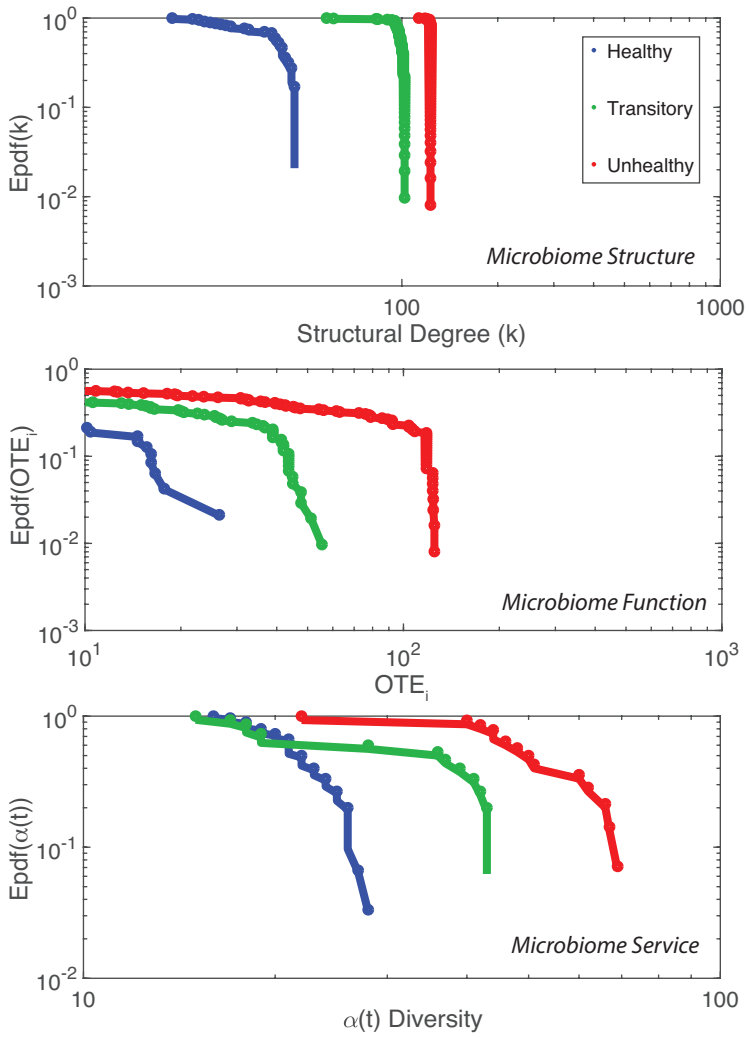

Figure 3.5: **Exceedance probability distribution of microbiome structure, function, and service.** Network degree, total outgoing transfer entropy (OTE) of each node, and α-diversity over time characterize the structure, function and service of the microbiome network.

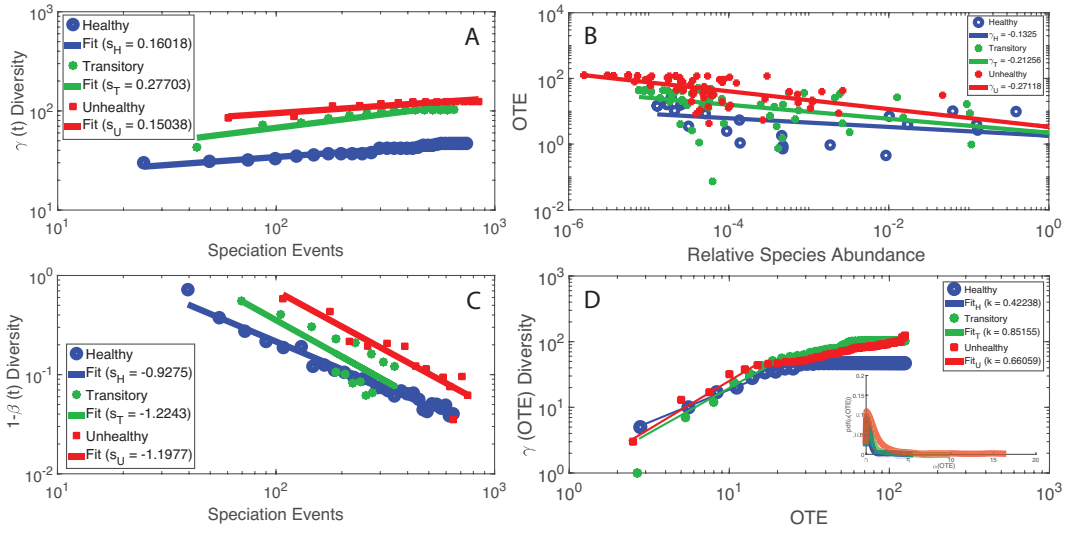

Figure 3.6: **Macroecological scaling patterns and predicted species interactions.** (**Left**) The scaling of total γ-diversity and species similarity $1 - \beta$ dependent on the number of speciation events (that is the number of new and existing species introduced until the time considered); speciation time is a proxy of the sampling area over time. (**Right**) The scaling of OTE vs. RSA and γ-diversity vs. OTE that consider the mutual variability of information exchange and macroecological indicators of the microbiome.

Chapter 4

Temperature-driven organization of fish ecosystems and fishery implications

4.1 Introduction

4.1.1 Impacts of Ocean Warming on Marine Fisheries

Marine fishes and invertebrates remain one of major sources of food, nutrition feeding hundreds of millions of people around the world, and have become of increasing importance, especially for coastal and developing countries, where they provide inhabitants with as much as 50% of animal protein intake [145]. According to annual reports from FAO, global food fish consumption grew at an average rate of 3.1% from 1960s, and the average food fish consumption per capita had increased from 9.0 Kg in 1961 to 20.5 Kg in 2018, with the rate of about 1.5% per year [146]. Despite their critical role in global food security and nutrition, fishery ecosystems are all along under increasing anthropogenic pressure from pollution, habitat degradation, overfishing and climate change that has been steadily warming the sea water. As of 2015, 33.1% of global fish stocks were assessed as fished at a rate faster than they can reproduce and therefore overexploited, while only 10% in 1974. Marine fishery ecosystems are complex systems that are highly dynamical and sensitive to external environmental factors including sea temperature. Ocean warming is altering fishery ecosystems and generally pushing negative impacts on the populations and behavior of fish species. Data from the US National Oceanic and At-

mospheric Administration (NOAA) shows that sea temperature anomaly increased from -0.02°C in 1974 to 0.79°C in 2016 which is the largest anomaly observed in last 100 years [147]. By comparing sea temperature data from last two decades (1999-2019) to two decades before that (1979-1999), it is observed that the rate of ocean warming had increased over 430%, meaning that ocean warming had dramatically sped up in last two decades [147, 148]. "Fish are like Goldilocks: They dont like their water too hot or too cold, but just right.", said Malin L. Pinsky, a co-author of the study [149], where they found that some few species populations benefited from ocean warming, but more of them suffered. Most fishes are not tolerant of abnormal fluctuations of sea temperature. From the perspective of fishery catches, ocean warming have reduced the sustainable harvest from a wide range of species by 4.1% (1.4 million metric tons of fish, approximately) since 1930s [149]. It has been making fish species diversity and abundance decline, and putting food supply at risk for millions of people around the world.

Due to increasing societal concerns about the sustainability of marine fish ecosystems and relevant ecological degradation under the conditions of ocean warming, assessing the vulnerability of these ecosystems to ocean warming and identifying what the sea temperature would impose on ecosystems have received considerable interest. Currently, most studies use purely biological models to conduct these research topics on temporal scale. Through collecting and visualizing macroecological time-series data, observing the evolution of species abundance, distribution and diversity over time, and comparing macroecological evolution to the variation of oceanographic climatic data, they concluded that marine fish ecosystems are deteriorating, and one of the reasons for the deterioration is the increasing fluctuation of sea temperature caused by climate change. For example, *Cheung, W. et al.* found that increasingly warming sea temperature altered the composition of marine fish catches, with an increase in the proportion of warmer water species catches at higher latitudes and a decrease of subtropical species catches at lower latitudes. They plotted sea surface temperature in the past four decades and calculated mean temperature as an indicator describing the preference temperature of fish species quantified as annual fish catches of 52 large marine ecosystems [150]. *Christopher M. Free et al.* applied temperature-dependent population models to measure effects of ocean warming on the productivity of 235 populations of 124 species in 38 ecological regions. They analyzed the data of marine fishery production from 1930 to 2010 and computed maximum sustainable yield over time. These results addressed that maximum sustainable yield of considered fish populations decreased

by 4.1% in five ecological regions experiencing losses of 15 to 35% during the studied period [149]. *Aidan Hunter et al.* used the time-series data of herring and sprat populations collected from North Sea since 1960s and west of Scotland since 1980s to identify long-term trend of species' growth rate and maturation schedules in average [151]. They visualized abundance and temperature, and estimated the trend of length at age and growth parameters including absolute growth rate over time. This study confirmed the fact that sea temperature is correlated to growth and maturation of fish species, and the mean length of herring populations steadily declined across multiple age groups. *Hubert du Pontavice et al.* examined influences of sea temperature on global biomass transfers from marine secondary production to fish stocks [152]. This study plotted the past observations of the trend of trophic transfer efficiency (TTE) and biomass residence time (BRT) in the seafood web from 1950 to 2010, and applied them to the projection of TTE and BRT by the end of the century. By visualizing the projection of TTE and BRT over time, they predicted that TTE and BRT would decrease until 2040 and remain relatively stable after 2040. All these studies, as examples, were conducted by using purely biological models on temporal scale and concluded biological results in particular fields in time domain. Results from these temporal biological models are important and valuable, but insufficient to completely describe marine fish ecosystems especially considering system dynamics and internal mechanism.

Scientists have recently paid great attention to investigating the impacts of ocean warming on marine fish ecosystems. Although conventional time-domain biological analyses are able to macroscopically display the change of fish species abundance, populations and diversity over sea temperature, it is hard to further understand how the fluctuation of sea temperature affects the collective behavior of marine fish communities, and information dynamics of ecosystems. In fact, macroscopic analyses on fish species are far from adequate to track the evolution of marine fish ecosystems under the conditions of ocean warming. In a particular fish community, there are a large number of fish species that play different roles including native and invasive species, prey and predator. In view of system, these species take various responsibilities for maintaining the marine fish ecosystem stable, and reshape the ecosystem in different ways. Therefore, investigating fish ecosystems simultaneously in temporal, systemic and species-specific sense is quite important and requisite to understand mutual relationships between fish species and sea temperature, and to find solutions to improve resilience of marine fish ecosystems to ocean warming. Purely biological analyses are incompetent to study internal responses of

fish communities to the fluctuation of sea temperature in terms of system dynamics, and to specifically identify the species that are most affected by sea temperature, and most responsible for system stability. In addition, methodologies that incorporate modern information technologies and computational techniques into modeling marine fish ecosystems and relevant prediction are still lacking.

4.1.2 Optimal Information Flow Model and Multi-scale Ecosystem Analysis

In this study, to overcome limitations of conventional time-domain biological analyses and disentangle the complexity of mutual relationships between marine fish ecosystems and sea temperature, we develop the optimal information flow (OIF) model and employ this information-theoretic model to study a fish community in Maizuru Bay that is managed by the Maizuru Fishery Research Station of Kyoto University [21]. OIF model is based on transfer entropy (TE) that provides an asymmetric approach to measure directed information flow between random variables (species) [23, 153]. It is used to infer interdependencies (defined as species interactions) between species in the fish community and reconstruct networks for describing the marine fish ecosystem, allowing us to study the ecosystem on both temporal and temperature-dependent scales in systemic way. OIF model is improved with respect to [31] by considering its extension over time to reconstruct dynamical information networks, the varied Markov order of random variables and a refined pattern-oriented criteria to select optimal time delay and threshold based on maximization of mutual information. Therefore, OIF overcomes the limited TE for the directed uncertainty reduction scheme, for the consideration of maximum information (entropy) network [30], and MI-based maximization criteria to define the optimal time delay and interaction threshold to accurately predict systems' patterns (e.g. biodiversity). It is noteworthy to mention that the optimal threshold on interactions is not necessarily within the scale-free or maximum entropy range of inferred collective behavior. In this way, interaction processes are clearly linked to patterns which provide relevance to the inference problem. Additionally, OIF is dependent on the choice of appropriate time delay between variables for TE calculation.

Considering the fish community, in order to specifically study the effects of sea temperature on the marine fish ecosystem in view of system, long-term abundance time-series data are categorized into five groups considering five temperature ranges (TR): $\leq$10°C, 10-15°C, 15-20°C, 20-25°C, $\geq$25°C. We detect mutual rela-

tionships between species and reconstruct networks for each TR group in attempt to investigate how the fluctuations of sea temperature affect the collective behavior of particular fish species and the whole fish ecosystem by conducting network-based analyses. For each TR group, mutual causal interactions between species are inferred by OIF model (computed as TE considering optimal time delay and threshold). These OIF-inferred species interactions define the connectivity among species, forming OIF networks for all TR groups. Structural and functional patterns are recognized for OIF networks. Network-based species-specific analyses are also conducted by analyzing how much a particular species interacts others, and identifying the most salient links and critical nodes in networks. Through comparing structural and functional patterns of and critical species in networks among five TR groups, we conclude the influences of fluctuations of sea temperature on the marine fish ecosystem and its stability and identify behavioral responses to sea temperature. In particular, temperature-dependently dynamical networks are also built and analyzed to tract the evolution of the marine fish ecosystem over sea temperature. As well, since time is the most common reference system in nature, temporally dynamical networks are also implemented to capture the change of marine fish ecosystem considering long-term ocean warming. Combining with temporal biological analyses, this work is expected to give a better understanding for marine fish ecosystems that help to maintain the ecosystems stable.

4.1.3 Stability, Sustainability and Management of Marine Fish Ecosystems

Oceans are home to a wondrous array of plants and animals including estimated 20,000 fish species, supporting the species bank of the planet, providing humans with food consumption, and maintaining the stability of global ecosystem and environment. Although it is impossible to know the exact number of species in the ocean (scientists estimate that more than 80% of oceanic species have yet to be discovered [154]), the number of species which people already knew in the ocean is decreasing. More importantly, the rate of species extinction is still increasing, making marine fish ecosystems in trouble under ocean warming caused by anthropogenic climate change. It is urgent and important to keep marine fish ecosystems stable, sustainable and well-managed for global environment and human health. People always understand ecosystem stability in terms of biodiversity, species richness and composition, and community structure. Studying ecosystem stability only consid-

ering these biomass and taxonomic indicators is not enough. Intuitively speaking, biodiversity increases ecosystem stability, but climate change or other human-driven perturbations may alter these positive relationships [155]. Therefore, these relationships has been contentious. In fact, the concept of ecosystem stability is complex due to multifaceted understanding and types [156–158]. As for a ecosystem, different types of stability describe different features that might lead to different patterns. Besides considering biomass and taxonomic indicators, ecosystem stability should be also explored from the perspective of the strength of species interactions, connective complex networks, topology of food webs, and resilience of species communities to different types of environmental perturbations [159].

In addition to biomass and taxonomic indicators from macroecological analyses, OIF is used to investigate ecosystem stability by inferring species interactions and reconstructing networks in this study. We macroscopically analyze the evolution of the marine fish ecosystem over sea temperature, and microscopically identify critical species who play a significant role in maintaining the structure and function of the fish ecosystem. In attempt to understand the dynamics of the fish ecosystem completely, results from biomass and taxonomic analysis and dynamical OIF interaction networks are combined together to study the fish community on multiple scales. This work would provide a potent tool to observe the stability of marine fish ecosystems, and be valuable to formulate science-based and accurate fishery policy to maintain these ecosystems stable and sustainable.

4.2 Methods

4.2.1 Time-series Data and Categorization

Long-term time-series data (in total 285 time points) of the fish community were collected by scientists at Maizuru Fisheries Research Station, Kyoto University [21]. They conducted underwater visual census approximately every two weeks along the coast of Maizuru Bay from 1 January 2002 to 2 April 2014 [160]. Such high-frequency time series enable the detection of interactions between species. Maizuru Bay is a typical semi-enclosed water area with nearly 50m of the shore and at a water depth of 0-10m, located in the west of Wakasa Bay, Japan. Precipitation is rather high from summer to winter which is the rainy season in this area. Sea surface temperature ranged from 5.2 to 31.8, sea bottom temperature from 8.5 to 29.6, which were measured near the surface and at the depth of 10m underwater

during the diving, respectively [161].

In this data set, only 14 dominant fish species whose total observation counts were higher than 1000, and 1 jellyfish species were included because rare species were not observed during most of census period. Rare species would bring large numbers of zeros in the time series which may lead to difficulty in analyzing data. Ignoring rare species does not significantly change the marine fish ecosystem and results of this work. Jellyfish, a non-fish species, was involved in the dataset since it was thought to have prominent influences on the dynamics of the fish community due to its large abundance. Note that time-series data were already normalized to unit mean and variance before analysis [21, 162].

In this study, we propose to study effects of water temperature on system dynamics, stability and sustainability of the fish community. To specifically do so, normalized time-series data were categorized into five subsets considering five mean temperature (the average of surface and bottom water temperature) ranges including $\leq$10°C, 10-15°C, 15-20°C, 20-25°C, $\geq$25°C, resulting in five shortened time series with 16, 93, 58, 62, 56 time points, and allowing us to separately analyze time series of these five subsets in order to disentangle how system dynamics of the fish community fluctuates with the change of sea temperature. It is worth noting that, as an assumption, the subset of time-series data after categorization in consideration of five temperature ranges are still regarded as sequential time series, called temperature-dependent time series here, even though the data categorization destroys the wholeness of original time series in terms of time order.

4.2.2 Probabilistic Portrayal of the Fish Community

We estimate the probability distribution of 1) the normalized abundance of fish species, and 2) inferred interactions between species (both are generally indicated as y as a generic random variable) in a way of using power-law distribution function [32, 124]. Theoretically, power-law function has two types: discrete and continuous, of which the continuous form is given by [163]:

$$p(y) = \frac{\varepsilon - 1}{y_{min}}(\frac{y}{y_{min}})^{-\varepsilon}, \tag{4.1}$$

where y_{min} is an estimated lower bound for which the power-law holds and $y_{min} > 0$. ε is the power-law scale exponent underlying the statistical behavior of variable studied. Power-law scale indicates how mean and variance of variables behave.

With respect to probabilistic characterization, distributions of data are visualized by computing discrete exceedance probability distribution (EPDF) that is defined as $P(Y \geq y) = 1 - P(Y < y)$, where $P(Y < y)$ is cumulative distribution function (CDF) [163] derived from probability distribution function $p(y)$ (pdf), and by plotting EPDF on log-log scale. We also introduce cutoff Y_{break} to the random variable whose probability distribution explicitly presents multiple regimes. These probability distribution regimes with different power-law scale exponents are separately estimated. Take EPDF with two regimes as an example, EPDF corresponding to this type of power-law distribution is formulated as [109]:

$$P(Y \geq y) \sim \begin{cases} (\frac{y}{y_{min}})^{-\varepsilon_1+1} f(y) & \text{for } y < Y_{break} \\ (\frac{y}{y_{min}})^{-\varepsilon_2+1} f(y) & \text{for } y \geq Y_{break} \end{cases}, \tag{4.2}$$

where Y_{break} is the break point isolating two regimes of power law distribution, $f(y)$ is a generality to formulate the cutoff (or homogeneity) function [109], ε_1 and ε_2 are power-law scale exponents of two regimes. This type of power law is called broken power-law function, which is a piecewise function consisting of two or more components with different power-law scales in combination with particular break points [110, 164].

4.2.3 Species Diversity and Abundance Characterization

To biologically observe the fish community as a local marine ecosystem, α diversity, one of the macroecological indicators, is introduced to describe the chronological fluctuation of species diversity, and how species diversity changes with the increase of temperature. Therefore, we utilize α diversity dependent on both dimensions: time and temperature, aimed at figuring out collective behavior and system dynamics of the fish community locally, temporally and spatially. α diversity referring to the biodiversity within a particular scale: period, area or ecosystem, for instance, is computed as the number of species in that scale. Given a set of unique species $\mathbf{S} = \{S_1, S_2, ..., S_n\}$ whose normalized abundances $\mathbf{X} = \{x_1, x_2, ..., x_n\}$ change over time, time-dependent α diversity $\alpha(t)$ is defined as in [31]:

$$\alpha(t) = \sum_{i=1}^{n} x_i(t)^0, \tag{4.3}$$

where $x_i(t)$ is the normalized abundance of species i at the time point t. For each time point, sea surface temperature and bottom temperature were recorded. We

average them as mean sea temperature and reorganize temporal α diversity considering the mean temperature corresponding to each time point, yielding temperature-dependent α diversity $\alpha(C)$.

Additionally, fish species in the fish community are distinguished into fish stocks (FS) [165, 166], native and invasive species groups as shown in Table C.1. We separately calculate the total abundance of these species groups indicated as EP (Ecological Productivity) [167], FS, native and invasive dependent on time and temperature using the formula:

$$A(T) = \sum_{i=1}^{n'} x_i(T), \tag{4.4}$$

where T can be time point or temperature point, $x_i(T)$ is the normalized abundance of species i at the time point T, or at the temperature point T, n' is the number of species in a particular subgroup (for all species group EP, n' is n equal to 15.).

4.2.4 Information-theoretical Pattern Recognition and Network Inference

Information-theoretical Diversity Indices and Uncertainty

Diversity indices including Shannon index ("Shannon-Wiener index") and Simpson's diversity index are statistical measures of species diversity in the fish community. They are not themselves diversities, but indices of diversity [63]. In completely ecological sense, real diversity is an unambiguous concept recording the number of species observed at a local scale and indicating species richness and evenness. It is essential for experimentalists to document biodiversity on Earth, and ultimate reflections of disturbance driven by human activities or climate change in a particular ecosystem. This is actually what raw data can to a restricted extent do, while it is patently inadequate to dig out internal mechanisms about how ecosystems response to perturbations in terms of system dynamics and collective behavior of species. Entropy and diversity indices have a wide variety of usages under different conditions according to particular need. Entropic diversity indices are capable of providing biological information on the mechanism of assembly including rarity, commonness and fluctuation of abundance of species in fish ecosystems than real taxonomic diversity. Therefore, we calculate Shannon entropy for each species in the fish community based on probabilistic estimation of species abundance. Shannon entropy of

species i is defined as $H(X_i) = -\sum_{m=1}^{v} p(x_i(m))log_2 p(x_i(m))$, where $p(x_i(m))$ is the probability of an event m of unique normalized abundance of species i, v is the number of all unique events of abundance. Considering the whole species community, taxonomic α diversity is defined as equation 4.3. It is obvious that the sum of Shannon entropy of all species $H_\alpha = \sum_{i=1}^{n} H(X_i)$ is proportional to Shannon diversity index of the fish community, which is an information-theoretical index [31], defined as $H = -\sum_{i=1}^{n} p_i log_2 p_i$ (where p_i is the probability of which the species i is observed in the community). Shannon diversity index assumes that all species are regarded as a sample reservoir and species observations are sampled randomly. These information-based indices form a potent tool to understand biological mechanisms and dynamics of fish ecosystems.

Optimal Information Flow Model

To investigate internal mechanism and collective behavior of the fish community, we estimate mutual causal interactions between fish species. Transfer Entropy (TE), a quantity in information theory coined in [23], is extensively used to measure information flow between two variables that is considered as one of the methods to quantify causal interactions. TE is mathematically defined as the amount of information that a source variable provides about the next state of a target variable in the context of the past of the target [23]. It provides a prediction- and probabilistic-based tool in detecting directed and dynamical causal interactions without demand for any prior knowledge or assuming any particular functional form to describe the mutual interactions among elements in a dynamical system. As a result, we understand the "causal interaction" as predictability that is easier to mathematically quantified in terms of directed uncertainty reduction. The calculation of TE between two variables builds on conditional probability and joint probability considering Markov order and historical values of these variables. It is denoted as $TE_{X_i \to X_j}(q, s, u)$ according to equation 2.4, where X_i and X_j stand for two random variables, q and s denote the Markov order of variables X_i and X_j, $x_i(t)$ and $x_j(t)$ are time-series observations, and u is the free varied source-target time delay that yields time lagged interactions. In this study, X_i and X_j are normalized abundances of species i and j in the fish community. TE assumes that all analyzed variables obey memoryless stochastic Markov chain process [168]. It indicates that future states of variable are only dependent on the current state, while not determined by states in the past. Thus, q and s are fixed as 1 under this assumption. TE calculation is sensitive to

data features including probability distribution, extreme values and zeros. In this study, we apply Kernel (model-free) and Gaussian model (parameter-free) to test the data set of the fish community [111] and compared the results with CCM model that is another well-documented algorithm to measure causal interaction between variables [20], and Pearson correlation coefficient (see Figure C.5). Figure C.5 shows that Kernel model provides higher resolution and gradient for TE estimation compared to Gaussian model. TE from Kernel presents similar pattern in the map (Figure C.5 C) to that of CCM model (Figure C.5 A) and Pearson correlation coefficient (Figure C.5 B). Therefore, Kernel model is selected as the TE estimator in the OIF model.

TE is commonly used as a powerful tool to estimate mutual interactions in non-linear ecosystem thanks to properties of inherent non-linearity, asymmetry, no model assumption (model-free) and predictability between variables [26, 27, 169, 170]. After computing TE between species, directed interaction network visualizing the interdependencies among all species is reconstructed from TE matrix. Networks are always hard to distinguish due to large numbers of connections. In order to reduce redundant information, we first select an appropriate time delay u that minimizes the statistical distance between species defined as equation 4.5 [59], equivalently maximizes the mutual information between two species.

$$d(X_i, X_j) = e^{-I(X_i;X_j)}, \tag{4.5}$$

where $I(X_i; X_j)$ is mutual information (MI) between species i and j. MI is given in information theory as:

$$I(X_i; X_j) = \sum_{x_j} \sum_{x_i} p(x_i(t), x_j(t)) log \frac{p(x_i(t), x_j(t))}{p(x_i(t))p(x_j(t))}, \tag{4.6}$$

where $p(x_i(t))$ and $p(x_j(t))$ are marginal distributions of species i and j and $p(x_i(t), x_j(t))$ is joint distribution of these two species. In this study, we specify the time delay u ranged from 0 to 10. According to equation 4.6, time delay u is selected as an integer in the range that maximizes the mutual information. It implies that u is determined by considering predictability rather than "true" causality which is strenuous to concretely estimate.

Afterwards, an appropriate value is chosen as a threshold to filter TE values. This is the second step to do the redundancy reduction by removing weak interac-

tions (links) in networks. This step is formulated as:

$$f(X_i \to X_j) = \begin{cases} TE_{X_i \to X_j} & \text{for } TE_{X_i \to X_j} \geq TE_{thre} \\ 0 & \text{for } TE_{X_i \to X_j} < TE_{thre} \end{cases}, \tag{4.7}$$

where $f(X_i \to X_j)$ is the quantification of species interactions from species i to j, TE_{thre} is the threshold chosen to filter TE values. In this study, TE_{thre} can be a fixed value 0.01 or top 20% TE value. This two-step scenario of redundancy reduction forms the proposed optimal information flow (OIF) model.

Assessment of Network Stability and Species Ecological Importance

By applying OIF model to infer species interdependence, TE matrix is obtained to quantify causal interactions between all pairs of species, and defines the structure of reconstructed networks. Firstly, we exploit eigenvalues of TE matrix as an indicator to evaluate network stability [21, 76, 171]. Let matrix $\boldsymbol{W}$ be the TE matrix, if there is a vector $\boldsymbol{v} \in \Re^n \neq \mathbf{0}$ that satisfies:

$$\boldsymbol{W}\boldsymbol{v} = \xi \boldsymbol{v}, \tag{4.8}$$

then scalar ξ is called the eigenvalue of $\boldsymbol{W}$ with corresponding right eigenvector $\boldsymbol{v}$. Eigenvalue is a nonzero value which can be real or complex. Real part of eigenvalues determines whether and how fast the network returns to equilibrium from a perturbation. Imaginary part of eigenvalues indicates the frequency of oscillation during the return to equilibrium.

Secondly, we evaluate ecological importance of species based on OIF network, and identify critical species in the fish community (indicated as nodes in OIF networks) using information-theoretic index and link salience measurement. The information-theoretic index is defined as total outgoing transfer entropy (OTE, computed as $OTE(i) = \sum_j TE_{i \to j}$) [31] that measures how much one species affects others totally. $OTE(i)$ is the total information transition from species i transmits to all other species. It can be therefore interpreted as how much one species helps to predict others in terms of predictability. OTE index is able to measure species influences with directions thanks to the asymmetric property of TE.

On the other hand, we also employ link salience that was introduced in [172] to measure the importance of links in OIF networks. Link salience approach is based on the concept of effective distance $dist(i, j)$ (computed as $dist(i, j) = 1/W_{ij}$). It is intuitively assumed that strongly (weakly) interacting nodes are close to (distant

from) each other. In our heterogeneous networks with real-valued weights, the algorithm of the shortest-path tree (SPT) that identifies the most efficient routes from a reference node r to the rest of the network is implemented for all nodes ($SPT(r)$). Then $SPT(r)$ is represented by a $n \times n$ matrix with element $spt_{ij}(r) = 1$ if the link (i, j) is involved in the collection of shortest paths, and $spt_{ij}(r) = 0$ if it is not. In conclusion, the link salience of OIF network is defined as:

$$SAL = \frac{1}{n} \sum_{r=1}^{n} SPT(r), \tag{4.9}$$

Therefore, SAL is a linear superposition of all $SPTs$. If the element $sal(i, j) = 0$, link (i, j) has no role in networks; if $sal(i, j) = 1$, link (i, j) is important for all reference nodes; and if $sal(i, j) = 1/2$, link (i, j) is important for only half of reference nodes [172]. Node (species) importance is quantified by counting the frequency that one node exists in the most salient links as the reference terminal node (species).

4.2.5 Temporally and Temperature-dependently Dynamical Network Analysis

We also apply OIF model to develop dynamical networks by truncating the whole time series respectively considering time and temperature. The total number of time-series units on which dynamical network inference is carried out by OIF model is $G = \lfloor \frac{L-l}{\Delta l} \rfloor + 1$, where $\lfloor \bullet \rfloor$ rounds G to the smaller integer. When reconstructing temporally dynamical networks, L is the total number of the whole time series (here, $L = 285$), l is the length of each time-series unit, Δl is the numerical inter-observation (or time step). In this study, each time period is set as one year (the length of time-series unit l is 24 time points), time step Δl is 2 time points that correspond to one month, the whole time series is therefore truncated into 131 time-series units in total. When reconstructing temperature-dependently dynamical networks considering temperature, L is the greater integer of the maximum mean temperature, l is the upper limit of the first mean temperature range, Δl is the temperature step. In this data set, maximum mean temperature is 30.7°C (thus, L is 31). The first mean temperature range is $[6, 10]$ that means l is 10, and Δl is 1. Therefore, the whole time series is truncated into 22 time-series units corresponding to different temperature ranges whose difference is 5 except for the first temperature

range $[6, 10]$. Note that the length of each time-series unit corresponding to each mean temperature range can be different.

After temporal and temperature-dependent reorganization, OIF model is used to infer species interaction (TE) matrix for each time-series unit, resulting in temporal or temperature-dependent interaction matrices $\boldsymbol{W}_{i,j}(g)$ and dynamical networks. Here, g is the time point of each time-series unit for temporally dynamical networks, while the lower limit of the temperature range for temperature-dependently dynamical networks. We analyze dynamical networks by computing total interaction (TI), dominant eigenvalues and estimated effective α diversity of each dynamical network. TI is defined as the sum of all TE values in interaction matrix $\boldsymbol{W}_{i,j}(g)$:

$$TI(g) = \sum_{i,j} \boldsymbol{W}_{i,j}(g). \tag{4.10}$$

Dynamical Stability (DS) is calculated as the absolute value of real part of dominant eigenvalue:

$$DS(g) = |Re(\xi_{max}(g))|, \tag{4.11}$$

where $\xi_{max}(g)$ is the dominant eigenvalue of an interaction matrix. Estimated effective α diversity $\alpha_e(g)$ is defined as the number of nodes (species) involved in a dynamical network which is formulated as:

$$\alpha_e(g) = \sum_{i=1}^{n} h_i(g), \tag{4.12}$$

where

$$h_i(g) = \begin{cases} 0, & \text{for } \sum_{j=1}^{n}(|\boldsymbol{W}_{i,j}(g)| + |\boldsymbol{W}_{j,i}(g)|) = 0 \\ 1, & \text{for } \sum_{j=1}^{n}(|\boldsymbol{W}_{i,j}(g)| + |\boldsymbol{W}_{j,i}(g)|) \neq 0 \end{cases}. \tag{4.13}$$

Therefore, $\alpha_e(g)$ denotes the total number of nodes whose structural degrees are not zero in a dynamical network.

4.3 Results

4.3.1 Temporal and Temperature-driven Biomass Analysis

A simple analysis for original data of fish species abundance starts by looking into temporal trajectories and seasonal fluctuations of sea water temperature (Figure

4.1A), taxonomic α diversity (Figure 4.1B) and normalized abundance (Figure 4.2). It is evident that sea surface and bottom temperature, α diversity of the fish community fluctuated over time synchronously and seasonally. Approximately from 2007, seasonal fluctuations of sea surface and bottom temperature were slightly getting higher except for 2009 with the least fluctuation (Figure 4.1A), while the global trend of α diversity decreased (Figure 4.1B) in the meanwhile. This result means that increasing fluctuations of sea temperature lead to biodiversity loss for the fish community. The global decrease of biodiversity can be interpreted as a biological response to increasing fluctuations of sea temperature. Total abundance of EP, FS, native and invasive groups over time (Figure 4.2A, B) shows a slight increase with fluctuations, that is opposite to the global trend of species diversity, from 2007 with the exception of a spike in 2009 and a downward spiral in 2011. The highest total abundance of species in EP, FS and native groups in 2009 is likely owing to the least temperature fluctuation in this year. This result implies that increasing sea temperature makes some species more abundant in a global view, even though it negatively affects species diversity, and that species diversity and richness in the fish community is sensitive to the instability of sea temperature, even a marginally high fluctuation. These results can be more explicitly confirmed in Figure 4.2C, D where total abundance of EP, FS, native, invasive groups and species 1 (*Aurelia.sp*) is observed to increase exponentially over the increasing sea temperature [173, 174]. Therein, the abundance of species 2 (*engraulis japonicus*), the one in FS group as an exception, decreases. Species whose abundance decreases with the increasing temperature are supposed to compete with other species and present disadvantage in species interactions. This competitive disadvantage may lead to departure and extinction of some species, resulting in the global decrease of species diversity in the fish community. We calculate α diversity and reorganize considering temperature, yielding α diversity against temperature shown in Figure 4.1C. It shows that α diversity grows with the increasing temperature, while the rate of the increase of α diversity ($\alpha'(C)$) gradually declines (see the black line fitting the $\alpha(C)$ points in the plot). The decreasing $\alpha'(C)$ implies that fish diversity does not continuously grow with the increase of temperature. Within lower temperature ranges, the fish community is more sensitive to the change of temperature relative to higher temperature ranges.

Time series of normalized abundance are categorized into five groups considering five mean temperature ranges (TRs): $\leq$10°C, 10-15°C, 15-20°C, 20-25°C, $\geq$25°C. Probability distributions of abundance data for these five TR groups are

characterized by plotting EPDF on doubly logarithmic axes and estimating exponential parameters of power-law fitting (Figure 4.3A) [109]. EPDF plots and decreasing power-law scaling (see the subplot in Figure 4.3A) show that the distribution of abundance becomes more scale-free with the increase of temperature. Lower scaling of power-law means fatter tail of power law distribution, indicating that the abundance of fish species is distributed more evenly for higher TR groups since some species become abundant. These results suggest that for higher temperature, fish species presents wider distribution in abundance compared to the lower temperature. Figure 4.3B shows standard deviation against mean abundance for each species corresponding to each TR group. Positive slopes of linear regression for these std-mean scatters address that fluctuations of species abundance scale with its magnitude. Exponents of scaling law for $\leq$10°C, 10-15°C, 15-20°C groups are higher than those of 20-25°C, $\geq$25°C groups (see the subplot in Figure 4.3B). This result confirms the finding from Figure 4.1C that the abundance of species in lower TR groups ($\leq$10°C, 10-15°C, 15-20°C) is more sensitive to the change of temperature compared to higher TR groups (20-25°C, $\geq$25°C). Sharp decrease of scaling law implies that the fish community experiences a significant change in collective behavior around the temperature of 20°C.

4.3.2 Interaction Inference and Temperature-dependent Network Characterization

Different patterns in EPDF and std-mean of species abundances shown in Figure 4.3 envision the discrepancy in species interactions and dynamics of ecosystem for different TR groups. In this study, species interaction is quantified by TE that is an information-theoretic variable measuring the amount of directed information flow and evaluating connectivity between species. Networks for the whole time series and five TR groups (see Figure 4.4) are therefore inferred by TE-based OIF model, and graphically visualized with *Gephi* [175]. To refine network structure, links of interactions lower than 0.01 are discarded by setting a threshold to filter TE. Considering the structure of networks, the size of nodes is proportional to Shannon entropy of fish species. The color of nodes linearly scales with the total outgoing transfer entropy (OTE) of species: the greater the OTE of a fish species is, the warmer the color of the node. The width and color of edges (links) are proportional to interaction between species computed as TE: the greater the TE is, the wider the link becomes and the warmer the link's color is. The arrow of links stands for the direction of

interaction (TE). Optimal Information Flow (OIF) networks present different structural and functional properties for particular TR fish groups. Network for the whole time series shown in Figure 4.4A provides a static overview of the fish community without considering the dynamical change of temperature or seasonality. It outlines causal relationships between species in the fish community, but is inadequate to tackle the dynamics of fish ecosystems driven by the fluctuation of temperature, and identify biological and behavioral responses. Therefore, in order to understand how temperature affects the fish community, networks for five TR groups ($\leq$10°C, 10-15°C, 15-20°C, 20-25°C and $\geq$25°C) are reconstructed and listed in Figure 4.4 from B to F. Here we first define network size considering both the number of nodes and the total amount of interactions between species. By looking at the structure of networks (Figure 4.4 B,C,D,E,F), it is obviously observed that network size is larger for the fish community within a higher temperature range. Nodes connected by links with warm colors are regarded as a cluster in which species are significantly affecting others or affected by others. Network of 15-20°C group shows a larger cluster with stronger interactions between species compared to other TR groups. Yet, the distribution of species interactions in this TR group presents more evenness considering the gradient of links' color in networks. This finding can be obviously observed in Figure 4.4B',C',D',E',F' describing the phase mapping of TE-based interaction matrices. TE matrix for 15-20°C group exhibits higher resolution and gradient in numeric. This result reveals that the network for 15-20°C group seems to be less scale-free versus other networks, and stands on a critical point where the fish ecosystem is experiencing a phase transition from one stable state to a metastable state. When considering specific species within different TR groups, except for the $\leq$10°C and 10-15°C groups in a globally stable state, species 6 ,7, 8, 9 are always the nodes with warm color (high OTE) (see Figure 4.4 A,C,D,E,F) in networks. Yet, interactions between these species are higher than others considering the phase mapping of TE matrices shown in Figure 4.4. This finding means that these species are core species in this fish community that are interacting strongly with each other, present quite different dynamics in behavior compared to other species, and play a significant role in maintaining the ecosystem stable. It is noteworthy to mention that all these species are native species in Maizuru bay (see Table. C.1).

We set a threshold as top 20% TE considering the Pareto Principle (also known as the 80/20 rule) which specifies that roughly 80 percent of the consequences come from 20 percent of the causes in many events [176]. TE threshold following this rule is generally higher than 0.01 so that it would further reduce network size, mak-

ing inferred networks briefer relative to the networks shown in Figure 4.4. OIF networks for the whole time series and five TR groups using the threshold of top 20% TE are reconstructed using the same regulation as Figure 4.4 and shown in Figure C.6. After the further reduction of network size, networks in Figure C.6 ignore more functional details, but are clearer to identify the structural difference between networks compared to Figure 4.4. Furthermore, it is possible to statistically analyze the structural degree, in- and out-degree of nodes for each network. In Figure C.6, nodes with warm color have more links with others than those with cool color. Statistical analysis for nodal degree is shown in Figure C.7. Structural degree, in- and out-degree are getting higher with the increase of temperature as a whole. This result implies that warmer conditions make the fish community more socially connected. Separately, pdf of structural degree (see Figure C.7) shows that structural degree of nodes increases with the increasing temperature, and favors bimodal distribution with different shapes for $\leq$10°C, 10-15°C, 15-20°C, 20-25°C groups, while roughly uniform distribution for $\geq$25°C group. For $\leq$10°C group, most structural degrees are distributed within the lower unimodal range from 1 to 3. With the increasing temperature, more and more values of structural degree are distributed in the higher unimodal range. However, for the highest TR ($\geq$25°C) group, the pdf of structural degree presents more evenness within the range from 0 to 10 (uniform distribution) compared to other groups. This result implies that structural degree does not increase continuously with sea temperature, and that the fish community would become less interacting when sea temperature is too high. Analogous features can be observed in the distribution of in-degree (Figure C.7 B) and out-degree (Figure C.7 C). For each TR group, OTE and average abundance are calculated for each species without considering any threshold for TE. Species OTE favors bimodal distribution with short range for $\leq$10°C and 10-15°C groups, and broad range for 15-20°C, 20-25°C and $\geq$25°C groups (see Figure C.8 A). This result suggests that overall effects of species on others grow with the increasing temperature. In a particular TR group, OTE exponentially scales with species abundance (see Figure C.8 B).

4.3.3 Interaction Spectrum, Phase Transition and System Stability

Probability distribution of TE-based interactions is characterized by calculating discrete exceedance probability, then fitted using power-law function (see Figure 4.5).

Among five TR groups, the greatest interaction occurs in the highest TR ($\geq$25°C) group. This result indicates that species interactions increase with sea temperature. The scaling of power-law distribution (computed as $1-\varphi$) is lower for $\leq$10°C, 10-15°C, $\geq$25°C groups and the first regime (black fitting) of 15-20°C group compared to 20-25°C group and the second regime (green fitting) of 15-20°C group. Lower scaling of power-law fitting implies that species interactions distribute more evenly compared to those groups whose scaling of power-law fitting is higher, and that the distribution of species interactions favors more scale-free in terms of the pattern of EPDF. It is concluded that the fish community within $\leq$10°C and 10-15°C lies on a globally stable state. Fish community on global stability is in the situation that the fish ecosystem is highly resistant to long-term perturbations (change of species composition, sea temperature, for instance) [177, 178], and interspecific and intraspecific interactions are not too strong [179]. With the increase of temperature, the fish community begins to get interacting gradually and two regimes appear in the distribution of species interactions for 15-20°C group. Therefore, the distribution of species interactions for 15-20°C group is divided into two sections by setting a break point as 0.3686 [110, 164]. The scaling of these two regimes is estimated using power-law fitting separately. The first regime whose power-law scaling is lower indicates that the fish community is still partly on a globally stable state, while the second regime with a higher scaling presents a less scale-free pattern implying a metastable state. This result suggests that for 15-20°C temperature range, the fish community stands on a critical point where the fish ecosystem is experiencing a phase transition from the global stability to metastability, and that the collective behavior of species is experiencing a significant change driven by sea temperature. The change of collective behavior can be reflected by the fluctuation of species diversity and abundance. For example, a significant drop (transition) of scaling law is observed in standard deviation against mean abundance as temperature rises from 15-20°C to 20-25°C (see Figure 4.3B). The power-law fitting for the second regime of 15-20°C group and 20-25°C group is the highest scaling among five TR groups. It suggests that major species interactions continue to increase as temperature rises, and large numbers of minor species interactions emerge simultaneously, leading to a distribution pattern that most interactions are clustered within a lower range. This distribution pattern of the fish ecosystem implies a less scale-free state compared to other TR groups. Fish ecosystems within 15-20°C and 20-25°C are therefore metastable (also see Figure 4.4D,E,D',E'). With the continuous increase of temperature ($\geq$25°C), the power-law scaling of the species interaction distribution shows

a significant decrease that is opposite to sea temperature. This result indicates that several major species interactions continue to rise, while minor interactions disappear, leading to a flatter distribution and more scale-free pattern compared to 15-20°C and 20-25°C groups. The fish ecosystem returns to another stable state after fluctuations caused by sea temperature. This finding confirms the fact that the fish community would not get interacting continuously with the increase of sea temperature, yet the activity of fish species would decrease when temperature is too high. This newly established stable state is interpreted as locally stable state which addresses the fish ecosystem only resistant to small short-lived disturbances (minor fluctuation of sea temperature, for instance) [180]. By connecting scaling exponents of the power-law fitting for TR groups over sea temperature shown in Figure 4.5F, the trend of power-law exponents also explicitly shows that the dynamical fish ecosystem is subjected to phase transition from global stability to metastability, and finally returns to local stability with the increase of temperature. A bifurcation occurs considering the division of species interaction distribution for 15-20°C group. This result reveals that, during the phase transition from global stability to local stability, the fish ecosystem stands at a critical point within 15-20°C where the fish community experiences a significant change in collective behavior. These changes result in different structural and functional features for the fish ecosystem within different temperature ranges graphically shown in OIF networks (see Figure 4.4).

Eigenvalues and eigenvectors provide essential information for reflecting the stability of dynamical systems [181]. Generally, differential equations are used to mathematically describe system dynamics based on variables in the particular system, and eigenvalues and eigenvectors can be used as a method to solve differential equations. The real part of eigenvalues determines whether the system is stable, and the imaginary part determines whether the oscillation is damped or not. If the real part of eigenvalues is negative, the system is stable and damped if the imaginary part is also negative, while undamped if the imaginary part is positive. If the real part of eigenvalues is positive, the system is unstable [181]. In this study, the eigenvalues of each interaction matrix corresponding to each TR group are scattered together in a complex plane (Figure 4.6). This method displays the position of all eigenvalues in the form of an ellipse [76, 171]. If only considering the eigenvalues with positive real part that may result in instability, Figure 4.6 shows one eigenvalue with the lowest positive real part for $\leq$10°C group, no eigenvalue with positive real part for 10-15°C group, two eigenvalues with positive real part for 15-20°C group, one eigenvalue with a positive real part for 20-25°C while lower than that for 15-

20°C group, and one eigenvalue with the largest positive real part for ≥25°C group. These results highlight that the fish ecosystem within low sea temperature ranges (≤10°C and 10-15°C) is stable, while becomes more interacting, leading to an unstable or metastable state as the sea temperature rises, especially for the 15-20°C group which has two eigenvalues with positive real part.

4.3.4 Network-based Biological Importance and Critical Species Identification

Biological importance and critical species identification are conducted by further analyzing the structural features of OIF networks and combining interaction spectrum with abundance spectrum. To this end, we first measure the salience for all links in OIF networks [172]. Link salience computation is based on the effective proximity d_{ij} defined by the strength of mutual interactions: $d_{ij} = 1/TE_{ij}$. It is intuitively assumed that strongly (weakly) interacting nodes are close to (distant from) each other. Link salience matrices shown in Figure 4.7 illustrate that species 7 (*Pseudolabrus.sieboldi*) is always the reference node of the most salient links with the exception of ≤10°C group where the reference node of the most salient links is species 15 (*Rudarius.ercodes*) instead of species 7. According to the website of *fishbase* (`http://www.fishbase.org/summary/Pseudolabrus-sieboldi.html`), species 7 (*Pseudolabrus.sieboldi*) is a native species in northwest pacific ocean and mainly distributed in Japan waters [182]. These results reveal that native species 7 has the most critical influence on other species in the fish community, and plays a vital role in maintaining structure and function of networks.

We also calculate Shannon entropy for species to measure the fluctuation and uncertainty of species abundance, and compute species' OTE as another nodal (species) property to quantify how much the species totally influences others in OIF networks. Then, species rankings considering top 5 Shannon entropy and top 5 OTE are listed and shown in Figure C.9. On the one hand, the species ranking of Shannon entropy (left figures in Figure C.9) shows that species 2 or (and) 5 always have the greatest Shannon entropy that represents the highest uncertainty in abundance for these two species. This finding also can be observed in the original time series of abundance plotted in Figure C.3. Species 2 and 5 have the highest fluctuation of abundance and the most diverse events in numeric. On the other hand, the species ranking of OTE (right figures in Figure C.9) shows that for all TR groups except the ≤10°C group, species 7 always has the greatest OTE, and species 6, 7, 8 and 9 are

are often involved in top 5 OTE ranking for 15-20°C, 20-25°C and ≥25°C groups. This finding means that these species have significant effects on other species in the fish community, of which species 7 is the greatest. This result is in accordance with the results of link salience matrices shown in Figure 4.7. In a word, through measuring link salience and OTE, species 6, 7, 8, 9 can be identified as the critical species in the fish community. Moreover, probability distribution functions of abundance for all species (see Figure C.2) show that normalized species abundances of these species are more evenly distributed within relatively small ranges (the least range is [0,45) for species 7) compared to other species with wide distributions and extreme values. It demonstrates that trajectories of species abundance for species 6, 7, 8 and 9 are remarkably divergent from other species. Species with wide distributions and extreme values present more uncertainty in abundance, that is, Shannon entropy of these species is higher than that of species 6, 7, 8, 9. Transfer entropy from source variables (species) with low uncertainty in abundance (species 6, 7, 8, 9, for instance) to target variables with high uncertainty (see left figures in Figure C.9) is supposed to be high. Considering species 6, 7, 8 and 9 themselves, within the small range of abundance distribution, relatively speaking, these species also present transparent divergences (dissimilarity) in the distribution of species abundance. Therefore, mutual interactions (transfer entropy) between these species are also high. These results address that the distribution of species abundance can be used as a proxy of interactions by comparing the similarity and divergence. Take species 2 (*Engraulis.japonicus*) as an extreme example, TEs from species 2 to others are expected to be low since original data of species 2 have many zeros or values close to zero that lead to lower uncertainty despite the asynchrony and divergence.

Considering the whole time series, rather than solely estimate the simple continuous pdf of abundance for each species (Figure C.2), we probabilistically characterize distributions of abundance by computing EPDF and using power-law fitting. The scaling exponent of power-law fitting is considered as another species property in terms of the distribution of taxonomic abundance (see Figure C.1). Thereafter, abundance-interaction phase space describing the relationships between power-law scaling and species OTE (power-law exponents vs. OTE for all species, indeed) is shown in Figure 4.8. For a particular species, the greater the exponent of power-law fitting for the EPDF is, the more the abundance values are distributed within a lower range; the higher the OTE is, the stronger the species interacts others. The phase space mapping shows a rough trend as a whole that the exponent of power-law fitting for the EPDF of species abundance is proportional to the species OTE with the

exceptions of species 3 and 6. This result suggests that species whose abundance has a wide distribution with a fat tail have relatively low total effects on other species, while species whose abundance is evenly distributed within a narrow range have high total effects on other species (species 7, 8, 9, 10 in the figure, for instance). These findings are also observed in the phase mapping of interaction matrix for the whole time series shown in Figure 4.4A'. It is intuitively speculated that species with more uniform distribution play a cooperative role in the fish community that is beneficial to the increase of abundance, but do not have strong effects on other species. In Figure C.10, OTE vs. mean abundance shows that for most species, effects on other fishes are proportional to the abundance of species (Figure C.10 A), while inversely proportional to the standard deviation of the abundance of species (Figure C.10 B). This result implies that native species with relatively uniform distribution of abundance play a significant role in the fish community.

Additionally, to identify which species are most affected by sea temperature, TE, Pearson correlation coefficient (indicated as cc) and ρ of CCM [20] are used as indicator to measure the relationship between species abundance and sea temperature (Table 4.1). This table shows that TE from sea temperature to species abundance for species 5, 6 and 7 is higher than other species, yet MI and ρ are overall higher for species 5, 6, 7, 8 and 9. Species 5, 6, 7, 8 and 9 are therefore considered as the species that are most affected by sea temperature. By looking into Figure C.4, the abundance of species 5, 6, and 7 exponentially scales with sea temperature more obviously than other species. The results mean that OIF-inferred TE outperforms other indicators for this purpose. TE from species abundance to sea temperature is also involved in the table and the TE values are very small. These are intuitively understandable that fish species in the ocean do not affect environmental factors [21].

4.3.5 Temporally and Temperature-dependently Dynamical Interactions and Stability

Considering both dimensions of time and temperature, temporally and temperature-dependently dynamical species interaction (TE) matrices are inferred from time series of each time and temperature unit, respectively. On the temporal scale, Figure C.11 A shows the real part of the dominant eigenvalue of TE matrix (blue line) and adjacency matrix (red line) underlying the structure of temporally dynamical networks. The real part of the dominant eigenvalue of TE matrix over time is lower

Species	**TE(T→s_i)**	**TE(s_i→T)**	**MI**	ρ
Aurelia.sp	0.0233	0.0046	0.3158	0.0951
Engraulis.japonicus	0.0008	0.0143	0.1323	0.2902
Plotosus.lineatus	0.0292	0.0017	0.1981	0.3101
Sebastes.inermis	0.0144	0.0313	0.4109	0.3178
Trachurus.japonicus	0.1712	0.0135	0.6203	0.6235
Girella.punctata	0.1065	0.0002	0.4721	0.3847
Pseudolabrus.sieboldi	0.0709	0.0444	0.6913	0.6902
Halichoeres.poecilopterus	0.0212	0.0073	0.7122	0.4469
Halichoeres.tenuispinnis	0.0072	0.0063	0.7065	0.4239
Chaenogobius.gulosus	0.0021	0.0093	0.1274	0.0634
Pterogobius.zonoleucus	0.0040	0.0011	0.2761	0.1195
Tridentiger.trigonocephalus	0.0104	0.0244	0.2635	0.3158
Siganus.fuscescens	0.0256	0.0101	0.2071	0.2944
Sphyraena.pinguis	0.0386	0.0049	0.1942	0.1992
Rudarius.ercodes	0.0229	0.0153	0.3894	0.2265

Table 4.1: Indices measuring the relationship between sea temperature and species abundance.

than that of adjacency matrix, yet the former presents an obvious seasonal fluctuation. In the first half-year period (approx. from winter to early summer), it is observed that the real part of the dominant eigenvalue increases over time and reaches a spike during this period. Then it decreases in the second half-year period, while increases again at the end of year. The increasing trend or the high level of the real part of the dominant eigenvalue always appears within the time period with higher changes (fluctuations) of sea temperature (from April to June or from October to December) in a year which probably corresponds to 15-20°C to 20-25°C temperature ranges. These results suggest that the fluctuation of species interactions becomes higher and fluctuates more frequently during the time period with higher fluctuations of sea temperature, and that the fish ecosystem tends to be metastable

as the fish community becomes active. For adjacency matrix, the high level of the real part of the eigenvalue appears synchronously with that of TE matrix, while it is hard to identify the seasonality due to lower fluctuations. Figure C.11 B shows the magnitude of total interactions over time that is computed as the sum of TE values in interaction matrices of temporally dynamical networks. Total interaction over time presents clear seasonal fluctuations that highly synchronize with the seasonality observed in eigenvalues of TE matrix shown in Figure C.11 A (blue line). This finding verifies the result that species interactions in the fish community increase during the time period with high fluctuations of sea temperature. Furthermore, species OTE is calculated for each temporally dynamical network, obtaining 256 OTEs for each species. Pdfs of species OTE show that critical species 6, 7, 8, 9 have higher effects on other species in the fish community compared to others (see Figure C.12).

On the scale of temperature, Figure C.11 E shows how the real part of the dominant eigenvalue of TE matrix (blue line) and adjacency matrix (red line) underlying the structure of temperature-dependently dynamical networks changes with the increase of temperature, respectively. Both curves show that the real part of the dominant eigenvalue rises with the increasing temperature as a whole, while that of TE matrix has more informative fluctuations with higher frequency relative to adjacency matrix. Figure C.11 F shows the magnitude of total interaction over temperature stemmed from TE matrices of temperature-dependently dynamical networks. It is clearly observed that total interactions analogously fluctuate with the real part of the dominant eigenvalue of TE matrix shown in Figure C.11 E, as well as to the estimated effective α diversity over temperature derived from dynamical TE-based interaction matrices after filtering by top 20% TE shown in Figure C.11 H. From the perspective of network-based models, these results indicate that quantitative interactions between species in the fish community are more sensitive to fluctuations of sea temperature compared to macroecological α diversity. The fluctuation of species interactions provides a potent strategy to monitor and predict biological collective behavior and system dynamics of the fish community, especially for internal intra- and inter-specific mechanisms and their responses to the change of environmental factors. The effective α diversity in temporal and temperature-dependent dimensions is shown in Figure C.11 C,D,G,H.

4.4 Discussion

Growth, reproduction and living habits of organisms in the ocean, as well as external environmental factors including sea temperature and daytime present a transparent seasonality. Therein, sea temperature is often perceived as a dominant ingredient that drives the biological behavior of organisms in ecosystems [183, 184], while also affected by human-induced climate change. For marine fish ecosystems, as narrated in the part of introduction, ocean warming has been pushing significant negative impacts on marine ecosystems and fish species. In this paper, we study the multispecies fish community in Maizuru bay by conducting biomass and taxonomic analyses on temporal scale first. These analyses capture the seasonal fluctuations of sea temperature, species diversity and abundance and conclude that the increasing fluctuations of sea temperature result in global biodiversity loss. Particularly, biomass and taxonomic analyses are also conducted on temperature scale considering five temperature ranges to investigate how sea temperature affects the fish community in biomass and taxonomy. Temperature-dependent analyses capture the difference in macroecological indicators of the fish community among five TR groups. In addition, same analyses for EP, FS, native and invasive groups and two species (species 1 and 2) are performed to identify cooperative or competitive interactions in the fish community. The result is helpful to explain the finding that species diversity loses even though the abundance of some fish species grows under ocean warming. Some valuable results are obtained from the analysis considering species abundance and diversity, while it is not sufficient to completely display the impacts of sea temperature on the fish community, and limited to describe internal mechanisms of how sea temperature affects the dynamics of the fish ecosystem.

To better understand biological responses of the fish community to the fluctuations of sea temperature, the proposed information-theoretic OIF is employed to study the fish community considering species interactions and system dynamics. In information theory that has attracted attention in complex networks research, entropy (information) is used to quantify the uncertainty of a variable and its computation is based on the distribution of observations. Transfer entropy is an asymmetric variable that measures directed relationships between two random variables by estimating the directed information flow between variables. In complex ecosystems, the concept of entropy is also used as indices for measuring diversity (Shannon-Wiener and Gini-Simpson diversity, for instance) [63]. The fundamental work of studying complex ecosystems is to detect complex interdependencies comprised of

a large number of potential interactions between all species. Therefore, the intuitive and heuristic notion in information theory for species interaction detection is transfer entropy. Therefore, OIF model is introduced and deployed in this study to detect potential species interactions that form information flow networks modeling the fish ecosystem. Networks inferred from OIF model illustrate differences in structure and function among five TR groups. These structural and functional differences imply different system states and dynamics the fish community presents within different temperature ranges. For the lowest temperature range, the fish community lies on a globally stable state, then changes the state to metastability in intermediate temperature ranges, and finally returns to a locally stable state. As well, different patterns are recognized for internal mechanisms of the fish ecosystem within different temperature ranges by probabilistically characterizing species interactions. These shifts of ecosystem state and dynamics explain how sea temperature substantially affects the fish community. In addition, eigenvalues of TE-based interaction matrix are computed as indices to asses the ecosystem stability of the fish community [76, 171]. The determination of eigenvalues and eigenvectors of a system is extremely important to many problems in physics and engineering including stability analysis, oscillations of vibrating systems, to name only a few. It provides three measures of stability in terms of species interactions: (i) whether the fish community will return to the previous state after a certain perturbation, (ii) how fast the return will occur, and (iii) what the interactome dynamics of the fish community look like during the process of the return.

It is important to note that different species may respond to the fluctuation of sea temperature in different ways. As one of the factors, these different behavioral responses improve the complexity of the fish ecosystem, and could be exponentially proportional to the number of species in the fish community. To further tackle the complexity of system dynamics and identify the roles and importance of some critical species under the conditions of ocean warming, species-specific analyses are conducted by doing network-based computation and combining the network-based results with species diversity and abundance information. Firstly, link salience estimation is used as network-based analyses to classify the salient links in complex networks based on a consensus estimate of all nodes [172]. Here, we also apply this algorithm to identify critical species in the fish community by finding the common species in the classified salient links. Native species 7 (*Pseudolabrus.sieboldi*) is the most frequently observed species in the salient links as a reference node. This result means that species 7 is likely to play a more sig-

nificant role in maintaining the fish ecosystem stable compared to other species. Secondly, OTE that counts the total effects of one species on others is calculated for each species as a nodal property in networks. TE and OTE are considered as the strong indicators of ecosystem dynamics for pairwise and nodal functional characterization since those variables are related to species interdependence in view of complex networks [31]. These information-theoretic indicators based on the interdependence between species rather than pure biological species values lead to more informative biological meaning and offer better characterization and understanding for networks and their dynamical evolution. From the perspective of uncertainty reduction and predictability, it can be also said that TE measures how much the source variable helps to predict the target. Therefore, OIF-inferred networks are interpreted as predictive interaction networks, and OTE is a variable measuring how much one species contributes to the prediction of next states of the fish ecosystem. From this framework, species 6, 7, 8 and 9 are identified as critical species often involved in top 5 OTE ranking, and species 7 has the greatest OTE for all TR groups, but the $\leq$10°C group. This finding coincides with the results from salient link classification. In consideration of biomass and taxonomic analyses, the abundance of species 6, 7, 8 and 9 presents remarkable divergence from other species (see Figure C.2 and Figure C.3). For two divergent variables, TE from source variable with lower uncertainty to the target with higher uncertainty is supposed to be high. By combining these results together, it addresses that the distribution of species abundance can be used as a proxy of interactions in terms of divergence, asynchronicity and diversity of events.

We also develop dynamical networks dependent on time and temperature. In consideration of the relationship between eigenvalues and system stability, species interactions and effective diversity indicators, network-based analyses including the real part of the dominant eigenvalue of TE matrices, total interaction of dynamical networks and estimated effective α diversity are implemented in this section. Dominant eigenvalues over time reveal that the fish ecosystem is less stable (metastable) during the period with higher fluctuations of sea temperature in one year. This result is in line with the metastable state of 10-15°C and 15-20°C groups concluded in section 3.3, since these two temperature ranges are more likely to appear in spring and autumn which have relatively high temperature fluctuations in one year. Dominant eigenvalues over temperature also suggest that in a certain temperature range ($\leq$18°C), the increasing temperature rapidly makes the fish ecosystem more interacting, but less stable. Total interaction of dynamical networks present equivalent

fluctuations to eigenvalues on both scales that explicitly imply the relationship between species interaction and system stability. Moreover, effective α diversity from OIF networks is able to roughly track the real taxonomic diversity. Therefore, dynamical network analyses provide a real-time and temperature-dependent monitor that could directly display the system states and species activeness in view of system dynamics. This is important and useful to understand how the fish ecosystem responds to the change of sea temperature especially in the conditions of ocean warming, and to keep the fish community well-managed and stable.

4.5 Conclusions

Temperature is one of the dominant environmental factors that drives marine fish ecosystems and collective behavior of fish species. Abnormal fluctuations of sea temperature caused by climate change have created significant challenges for fish ecosystems [185]. In this study, the fish community in Maizuru Bay is studied in the time domain, as well as on temperature scale by considering biomass and taxonomy, species interactions and complex systems, and network-based species-specific identification.

- Time-domain analyses of biomass and taxonomy present seasonal fluctuations of sea temperature and α diversity. Slightly increasing fluctuations of sea temperature are observed in the figure of temperature vs. year, while the global level of species diversity decreases. Accordingly, the increasing fluctuations of sea temperature may lead to the biodiversity loss in the fish community. Total abundance of species in EP, FS, native and invasive groups slightly increases over year, and exponentially increases with the increase of sea temperature considering temperature scale. These findings indicate that increasing sea temperature leads to the rise of species abundance, even though it brings biodiversity loss in a global view. Biomass and taxonomic analyses for five TR groups show that species abundance is distributed more evenly for higher TR group, and the variance of species abundance scales with its magnitude. On temperature scale, taxonomic α diversity grows with the increasing temperature, and is more sensitive to the change of temperature for lower TR groups.

- We introduce OIF model to infer species interactions in the fish community, and reconstruct OIF networks for the whole time series and five TR groups.

OIF network for 15-20°C group seems to be less scale-free vs. other networks, indicating that the fish ecosystem within this temperature range is on critical metastable state, and experiencing phase transition from local stability ($\leq$10°C group) to metastability. In addition, collective behavior of species undergoes a significant change. From $\leq$10°C to $\geq$25°C group, nodal degree increases, but not consecutively when sea temperature continue to rise. High sea temperature makes the fish community socially connected, while would return to be silent when sea temperature is too high.

- Network-based analyses for species-specific identification based on the algorithm to classify salient links and OTE framework reveal that species 7 (*Pseudolabrus.sieboldi*) is likely to have the greatest influence on other species in the fish community and play a more important role in maintaining the fish ecosystem stable. OTE framework also recognizes a cluster involving species 6, 7, 8 and 9 who strongly interact each other. Note that species abundance of these species is remarkably divergent from others (see Figure C.2 and Figure C.3). It is concluded that the distribution of species abundance can be a proxy of species interaction in terms of divergence, asynchronicity and diversity of events.

- Temporally dynamical networks present the seasonality of stability. During the time period with high fluctuations of temperature (spring or autumn, for instance) in one year, the real part of the dominant eigenvalue of TE-based interaction matrix is high, suggesting that the fish community is more interacting and becomes metastable. Considering temperature-dependently dynamical networks, the real part of the dominant eigenvalue over temperature increases in a certain temperature range ($\leq$18°C). The increasing temperature makes the fish ecosystem more interacting, but less stable. The fluctuation of total interaction of dynamical networks is highly synchronous with that of the real part of the dominant eigenvalue. Effective α diversity from dynamical OIF networks is able to roughly track the taxonomic α diversity. Species interaction is more sensitive to the fluctuation of sea temperature compared to effective and taxonomic α diversity.

This study provides a two-dimensional (time and temperature) analysis for studying the fish community considering biomass and taxonomy, and system dynamics, respectively. Conventional time-domain analysis for biomass and taxonomy high-

lights the seasonal fluctuation of alpha diversity and species abundance. OIF model allows to investigate the fish community in view of species interactions and system dynamics. By reconstructing OIF networks, we study the evolution of ecosystem stability by analyzing structural and functional features of these networks. Network-based species-specific analyses are also conducted to identify critical species and clusters in the fish community, which might be most affected by the fluctuation of sea temperature, and responsible for keeping the fish ecosystem stable. Therefore, this work is important and useful to completely understand marine fish ecosystems and their responses to the fluctuation of sea temperature. It would be helpful to formulate science-based and accurate fishery policy to protect marine ecosystems and improve system resilience to ocean warming caused by climate change.

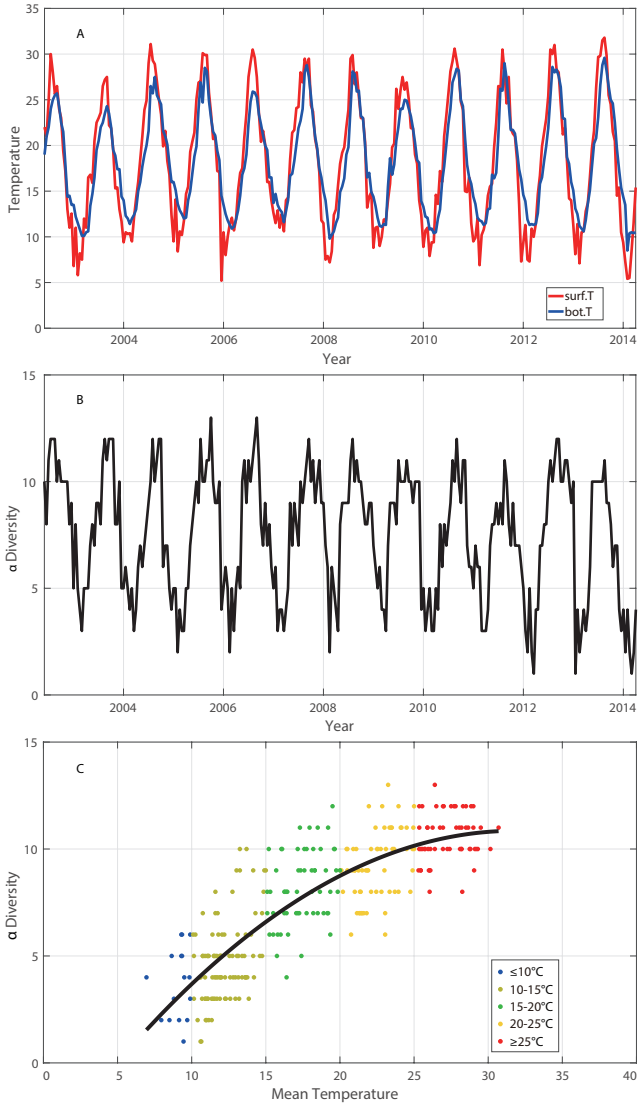

Figure 4.1: **Seasonal fluctuations of sea temperature and α diversity, α diversity over mean temperature. A**: The sea surface temperature (red line) and bottom temperature (blue line) over time within the range from June 2002 to April 2014. **B**: Taxonomic α diversity over time. **C**: α diversity over mean temperature (the average of sea surface and bottom temperature). Blue points, light green points, green points, yellow points and red points represent values of α diversity corresponding to different temperature ranges: $\leq$10°C, 10-15°C, 15-20°C, 20-25°C, $\geq$25°C, respectively. Black curve in plot C is a second degree polynomial fitting for α diversity over temperature.

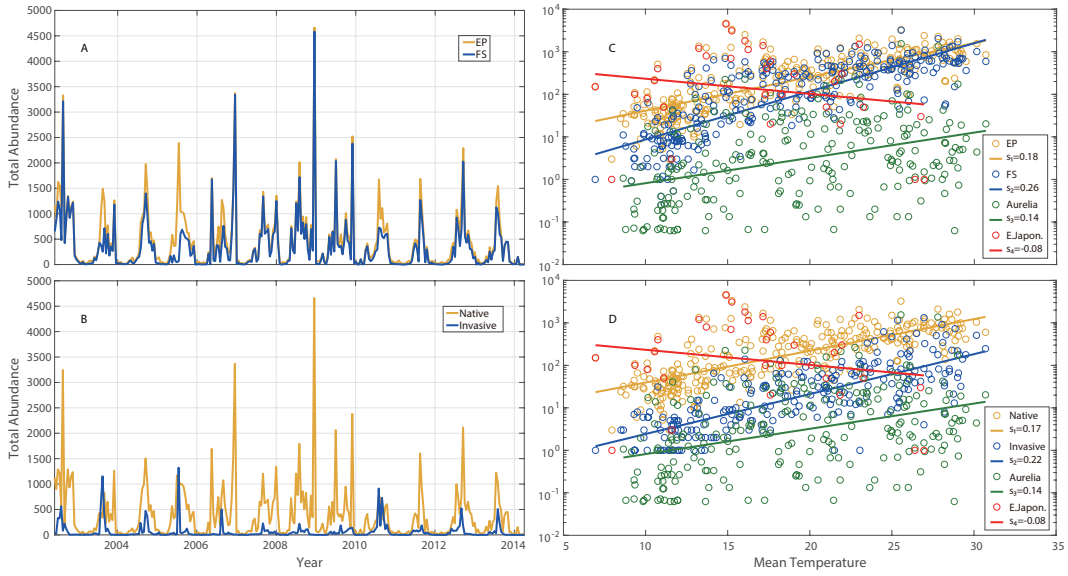

Figure 4.2: **Total species abundance of EP, FS, Native and Invasive groups over time and mean temperature.** **A**: Exceedance probability distribution function (EPDF) are scattered on log-log scale, and fitted by power-law function. $|-\epsilon+1|$ is the exponent of the original power-law function. All exponents for five TR groups are connected by a black dashed line shown in the subplot inside **A**. **B**: standard deviation against mean species abundance is plotted on log-log scale for five TR groups and fitted by power-law function. Here, v is the absolute slope (exponent) of power-law function. All exponents of the scaling law for five TR groups are connected by a black dashed line shown in the subplot inside **B**.

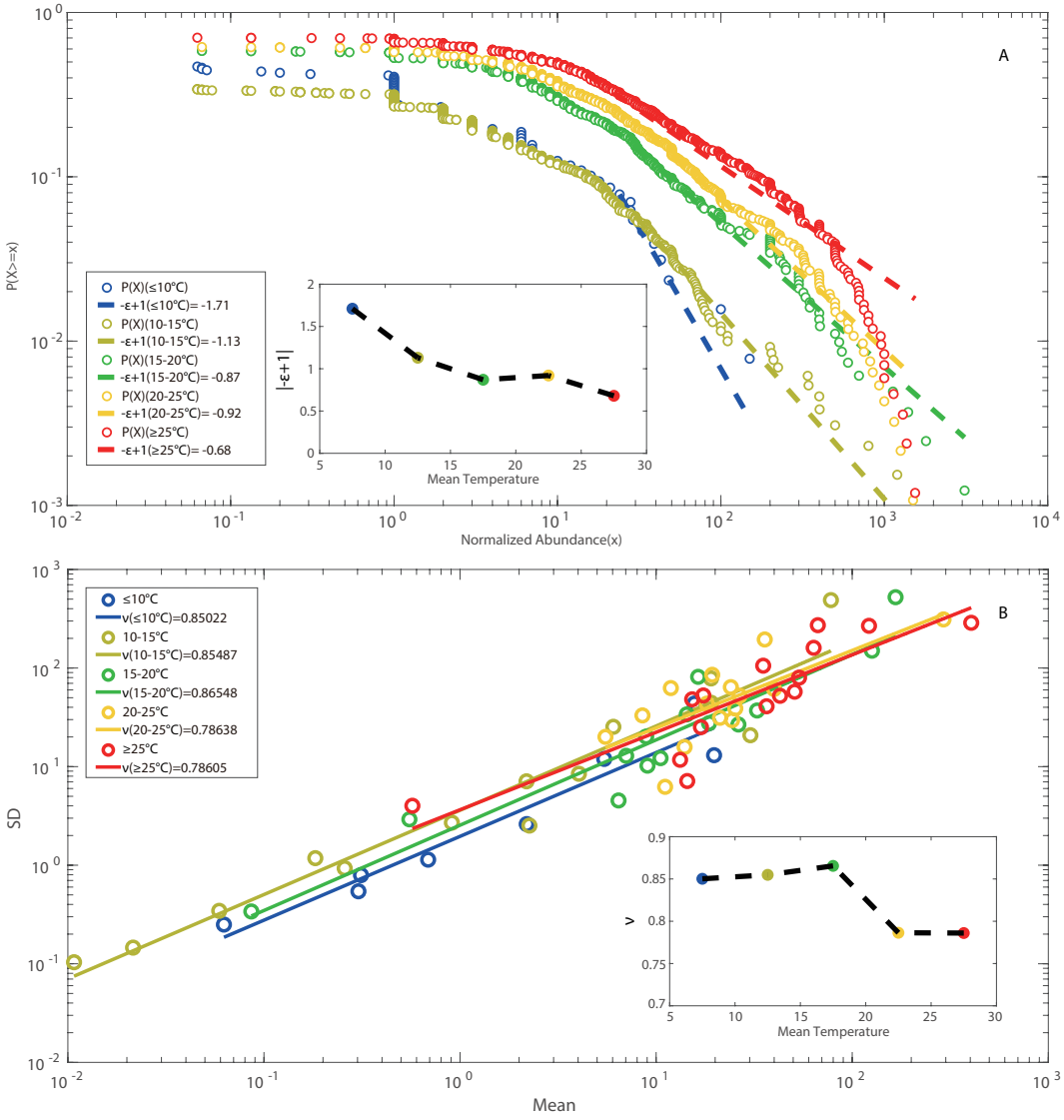

Figure 4.3: **Exceedance probability distribution, standard deviation vs. mean of species abundance. A**: Exceedance probability distribution function (EPDF) are scattered on log-log scale, and fitted by power-law function. $|-\epsilon+1|$ is the exponent of the original power-law function. All exponents for five TR groups are connected by a black dashed line shown in the subplot inside **A**. **B**: standard deviation against mean species abundance is plotted on log-log scale for five TR groups and fitted by power-law function. Here, v is the absolute slope (exponent) of power-law function. All exponents of the scaling law for five TR groups are connected by a black dashed line shown in the subplot inside **B**.

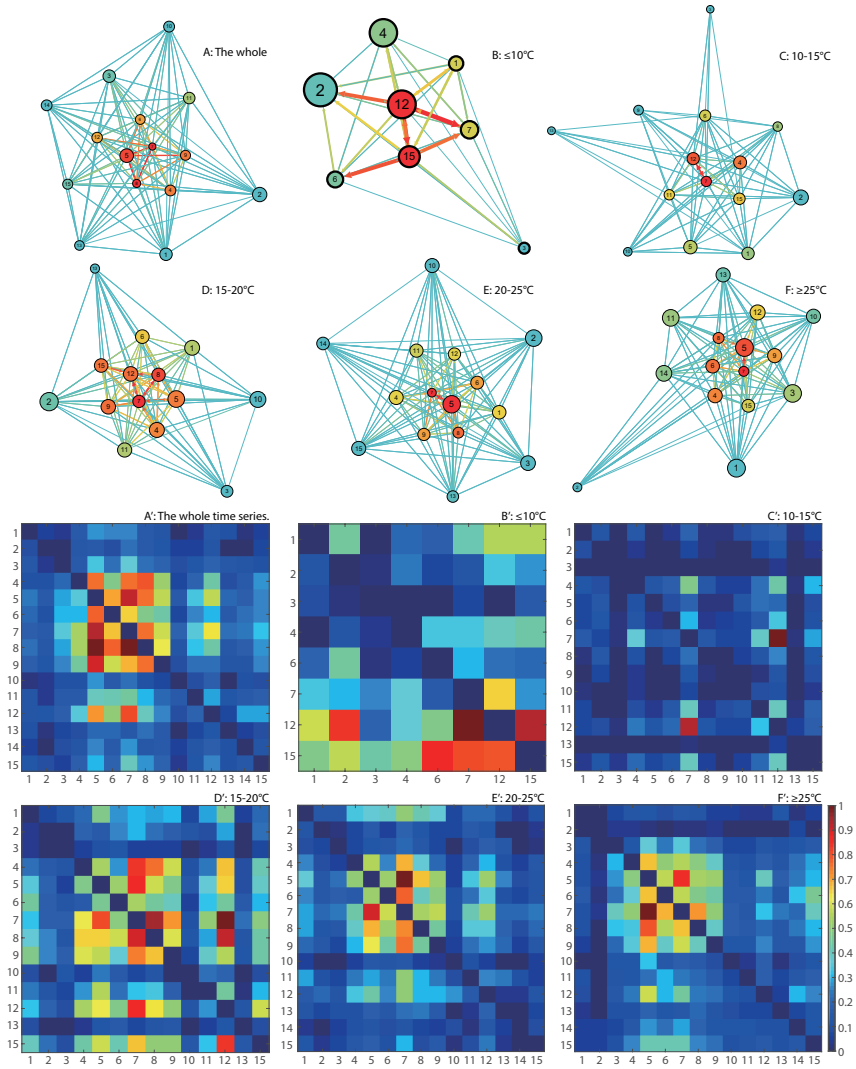

Figure 4.4: **OIF-inferred species interaction networks and matrices.** OIF model is used to infer the causal interaction between all pairs of species, yielding interaction (TE) matrices for the whole time series and five TR groups. TE values in interaction matrix are normalized to 1 and drawn in plots A', B', C', D', E' and F'. After removing weak interactions by setting a threshold (0.01) to filter transfer entropy, species interaction networks are reconstructed using *Gephi* and shown in plots A, B, C, D, E and F. The size of node is proportional to the Shannon Entropy of species, the color of node is proportional to the total outgoing transfer entropy (OTE) of species (the higher the OTE is, the warmer the node's color is.); the width and color of the link between species are proportional to the TE between the pair of species (The higher the TE is, the warmer (wider) the link's color (width) is.).

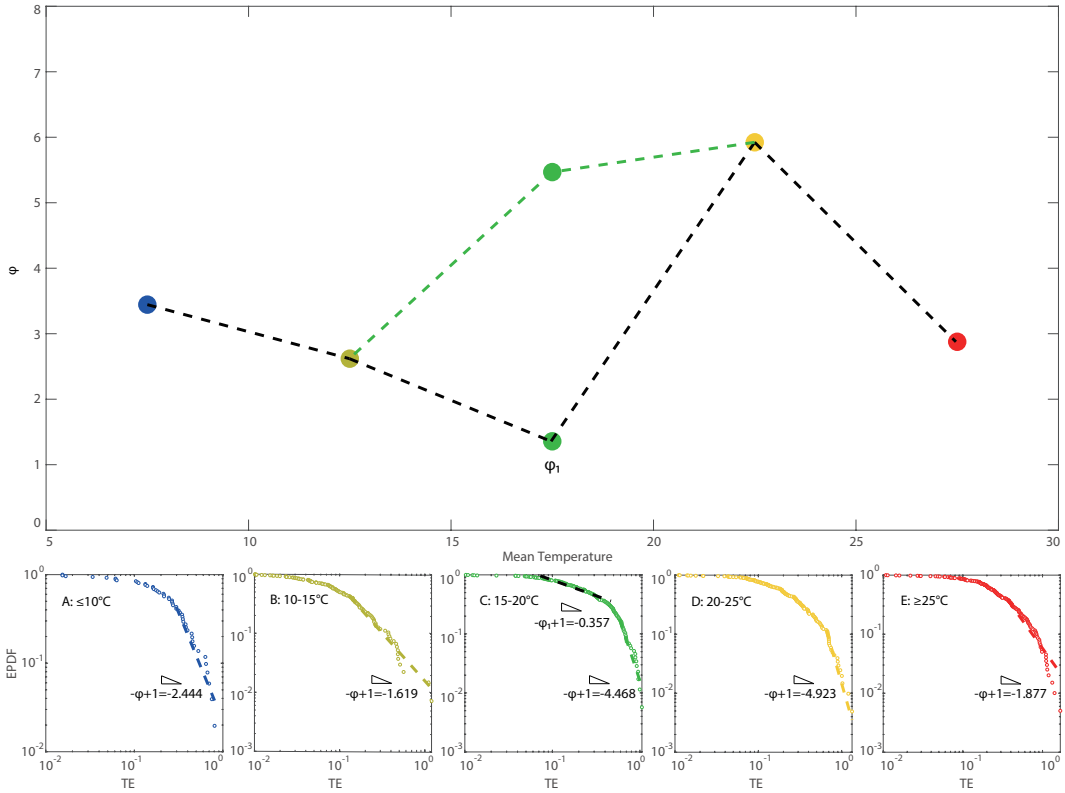

Figure 4.5: **Exceedance probability distribution function of species interactions after filtering with the threshold 0.01.** **A**, **B**, **C**, **D** and **E**: EPDFs of TE-based interactions between all pairs of species for five TRs: $\leq$10°C, 10-15°C, 15-20°C, 20-25°C, $\geq$25°C. They are shown on log-log scale and fitted by power-law function. In plot **F**, all exponents of the power-law fitting are connected by dashed lines. Scaling exponents for different TR groups are displayed in a coordinate plane where x-coordinates are middle values of temperature ranges (note that 7.5°C and 27.5°C are selected as middle temperature values for the range of temperature lower than 10°C, and higher than 25°C, respectively.), and connected by dashed line from low to high temperature. Note that for the 15-20°C group, EPDF of species interactions presents two regimes that are separately fitted by power-law function. The shape of the connection of exponents therefore presents a bifurcation within the temperature range of 15-20°C.

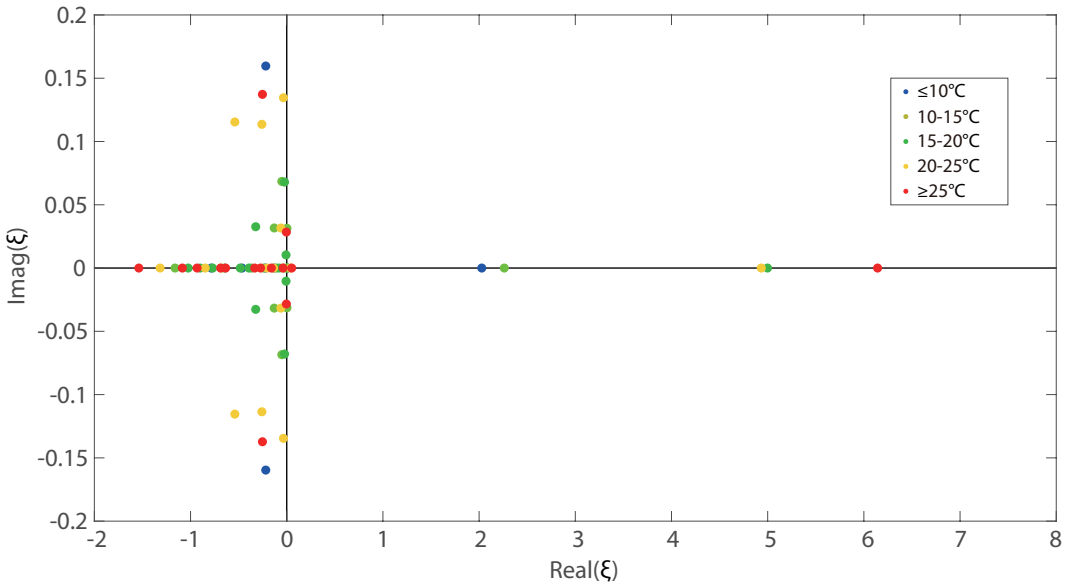

Figure 4.6: **Distribution of the eigenvalues of TE interaction matrices.** Eigenvalues of TE interaction matrices are scattered in a complex plane (five different colors correspond to five TR groups). Dynamical stability of the fish community is computed as the real part of the dominant eigenvalue of TE interaction matrices.

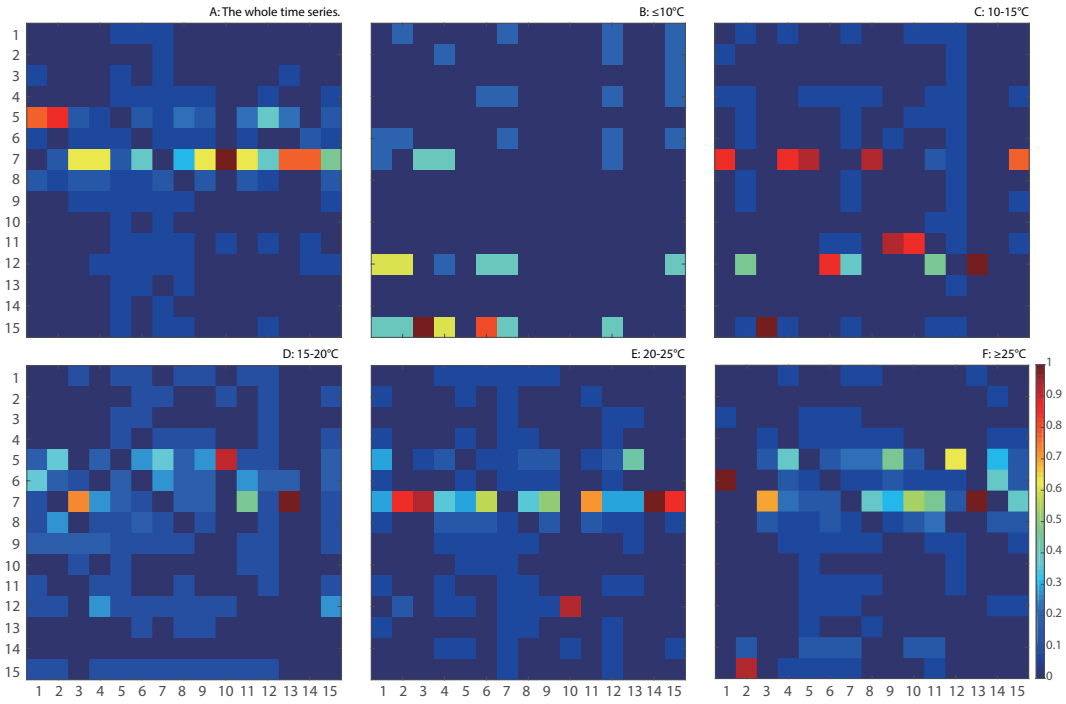

Figure 4.7: **Link salience matrices.** Link importance is measured by the algorithm of link salience for all OIF networks corresponding to the whole time series and five TR groups , yielding 15×15 matrices. The values of link salience are normalized to 1, and drawn in plots A, B, C, D, E and F, respectively.

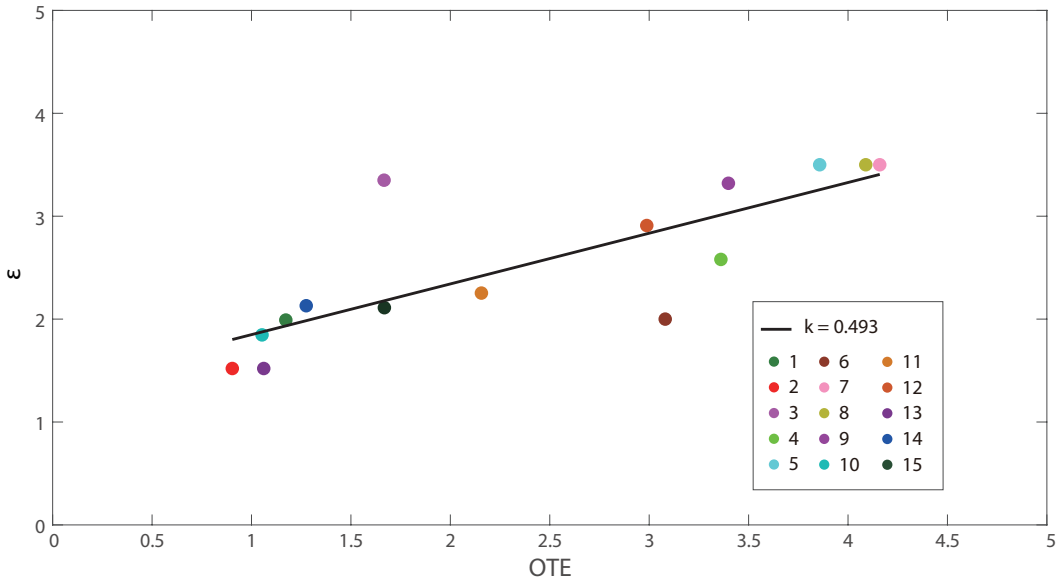

Figure 4.8: **The relationship between the distribution of species abundance and influences.** Considering the whole time series of 15 species, ϵ is the slope of the power-law fitting for the EPDF of species abundance (see Figure C.1), TE-base interactions are inferred by OIF model. OTE is interpreted as the total influences of one species on others. Black line is the linear fitting for ϵ vs. OTE of all species.

Chapter 5

Conclusions

The aim of this dissertation is to provide insight into the use of information-theoretic complex network inference model to untangle ecosystem complexity and information dynamics. An integrated Optimal Information Flow (OIF) model building on variables in information theory including Shannon entropy, mutual information and transfer entropy is introduced to quantify information flow (causal interactions) between all pairs of components and characterize the information dynamics of ecosystems. This OIF first provides a methodological framework for structural and functional quantification and visualization in view of information theory, leading to predictive species interaction networks. Through analyzing the information dynamics, structural and functional features of OIF-inferred networks and their alterations with environmental factors, OIF allows to identify the internal mechanisms of collective behavior as biological responses to environmental changes, and track the dynamical evolution of ecosystems. These results can be used to understand how a particular ecosystem responds to the fluctuations of the external environment with the help of macroecological analysis. Additionally, OIF links species interactions to taxonomic abundance and diversity. Species interaction can be a proxy of species abundance and diversity by comparing the similarity and divergence of time series. Therefore, this research provides a better understanding for ecosystems that is useful to improve the resilience to environmental disturbances, and create pathways to regulate ecosystems in due course under the pressure of external stressors.

Chapter 2 places the validation work for the developed OIF model. We investigate the ability of OIF model to infer bidirectional causality by comparing that to the well-documented CCM. The results from synthetic data generated by a simple predator-prey model, real-world data of a sardine-anchovy-temperature system and

of a multispecies fish ecosystem highlight that the developed OIF performs better than CCM to predict the patterns of species abundance and diversity. Specifically, OIF provides a more accurate inference for causal interactions with a larger gradient and smaller fluctuations, and higher accuracy in predicting α-diversity considering optimal time delays. Furthermore, an appropriate threshold on inferred causal interactions is proposed to maximize the accuracy in predicting the fluctuations of effective α-diversity, defined as the count of model-inferred interacting species in dynamical networks. Overall OIF outperforms other models in assessing predictive causality (also in terms of computational complexity) due to the explicit consideration of synchronization, divergence and diversity of events that define model sensitivity, uncertainty and complexity. Thus, OIF offers a broad ecological information by extracting predictive causal networks of complex ecosystems from time-series data in the space-time continuum. The accurate inference of species interactions at any biological scale of organization is highly valuable because it allows to predict biodiversity changes, for instance as a function of climate and other anthropogenic stressors. This has practical implications for defining optimal ecosystem management and design, such as fish stock prioritization and delineation of marine protected areas based on derived collective multispecies assembly. OIF can be applied to any complex systems and used for model evaluation and design where causality is considered as non-linear predictability of diverse events of populations or communities.

Chapter 3 explores the application of OIF model in gut-associated microbial ecosystems for the Irritable Bowel Syndrome (IBS). Based on OIF model with an threshold that maximizes the information content of inferred networks, we detect species interaction networks that are functionally and structurally different for healthy and unhealthy individuals. Healthy networks are characterized by a neutral symmetrical pattern of species interactions and scale-free topology versus random unhealthy networks. We also identify an inverse scaling relationship between species total outgoing information flow, meaningful of node interactivity, and relative species abundance (RSA). The top ten interacting species are also the least relatively abundant for the healthy microbiome and the most detrimental. These findings support the idea about the diminishing role of network hubs and how these should be defined considering the total outgoing information flow rather than the node degree. Macroecologically, the healthy microbiome is characterized by the highest Pareto total species diversity growth rate, the lowest species turnover, and the smallest variability of RSA for all species. This result challenges current views

that posit a universal association between healthy states and the highest absolute species diversity in ecosystems. Additionally, we show how the transitory microbiome is unstable and microbiome criticality is not necessarily at the phase transition between healthy and unhealthy states. We stress the importance of considering portfolios of interacting pairs versus single node dynamics when characterizing the microbiome and of ranking these pairs in terms of their interactions (i.e., species collective behavior) that shape transition from healthy to unhealthy states. The macroecological characterization of the microbiome is useful for public health and disease diagnosis and etiognosis, while species-specific analyses can detect beneficial species leading to personalized design of pre- and probiotic treatments and microbiome engineering.

Chapter 4 applies the OIF model to a multispecies fish ecosystem under ocean warming caused by climate change for understanding how the fluctuations of sea temperature affect the collective behavior and information dynamics of the fish ecosystem. Macroecological analysis addresses that the global increase of temperature from 2002 to 2014 reduces fish diversity, while some species become more abundant and that causes ecological productivity to grow exponentially. Specially to recognize the impacts of sea temperature on the fish community, the long-term time-series data are analyzed considering five temperature ranges: $\leq$10°C, 10-15°C, 15-20°C, 20-25°C, $\geq$25°C. OIF model is used to detect bidirectional interactions between species and reconstruct species interaction networks that are functionally different for each temperature range. Networks for lower and higher temperature ranges are more scale-free compared to networks for the intermediate 15-20°C range in which the fish ecosystem experiences a first order phase transition from a locally stable state to a metastable state. Species-specific analysis is conducted by calculating the link salience and total outgoing information flow. Native species whose abundance is distributed more uniformly have a higher total outgoing information flow, and are always the reference species (nodes in networks) of the most salient links (i.e. species 7). These species play an important role in maintaining the fish ecosystem stability and sustainability. In addition, It is observed that species diversity, total interactions and Shannon entropy of species abundance in the fish community grow with the increase of temperature. This work provides a data-driven tool for analyzing and monitoring fish ecosystems under the pressure of global warming or other stressors. Macroecological and network-based analyses are useful to formulate science-based and accurate fishery policy to maintain marine fish ecosystems stable and sustainable.

The results from all these studies evidence the efficiency of the developed OIF model. In particular, applications in real-world ecosystems highlight the strong performance and advantages of OIF in understanding collective behavior, information dynamics and system stability and evolution, and predicting biodiversity and species abundance in complex ecosystems. More importantly, OIF is competent to infer and predict system patterns and information dynamics not only in ecosystems studied in this work, but in many other complex systems (for instance, power grid systems, 5G networks, social networks and brain sciences) where traditional thinking and methodology may break down.

Bibliography

[1] M. Cadenasso, S. Pickett, and J. Grove, "Dimensions of ecosystem complexity: heterogeneity, connectivity, and history," *Ecological complexity*, vol. 3, no. 1, pp. 1–12, 2006.

[2] O. J. Schmitz, *Resolving ecosystem complexity (MPB-47)*. Princeton University Press, 2010, vol. 47.

[3] R. Solé and B. Goodwin, "How complexity pervades biology," *Basic Books*, 2000.

[4] L. Parrott, "Measuring ecological complexity," *Ecological Indicators*, vol. 10, no. 6, pp. 1069–1076, 2010.

[5] ——, "Complexity and the limits of ecological engineering," *Transactions of the ASAE*, vol. 45, no. 5, p. 1697, 2002.

[6] R. D. Holt, "Bringing the hutchinsonian niche into the 21st century: ecological and evolutionary perspectives," *Proceedings of the National Academy of Sciences*, vol. 106, no. Supplement 2, pp. 19 659–19 665, 2009.

[7] P. J. Morin, *Community ecology*. John Wiley & Sons, 2009.

[8] M. Vellend, "Conceptual synthesis in community ecology," *The Quarterly review of biology*, vol. 85, no. 2, pp. 183–206, 2010.

[9] V. Devictor, J. Clavel, R. Julliard, S. Lavergne, D. Mouillot, W. Thuiller, P. Venail, S. Villeger, and N. Mouquet, "Defining and measuring ecological specialization," *Journal of Applied Ecology*, vol. 47, no. 1, pp. 15–25, 2010.

[10] E. Delmas, M. Besson, M.-H. Brice, L. A. Burkle, G. V. Dalla Riva, M.-J. Fortin, D. Gravel, P. R. Guimarães Jr, D. H. Hembry, E. A. Newman *et al.*,

"Analysing ecological networks of species interactions," *Biological Reviews*, vol. 94, no. 1, pp. 16–36, 2019.

[11] P. Landi, H. O. Minoarivelo, Å. Brännström, C. Hui, and U. Dieckmann, "Complexity and stability of ecological networks: a review of the theory," *Population Ecology*, vol. 60, no. 4, pp. 319–345, 2018.

[12] L. Parrott and W. S. Meyer, "Future landscapes: managing within complexity," *Frontiers in Ecology and the Environment*, vol. 10, no. 7, pp. 382–389, 2012.

[13] T. Poisot, E. Canard, D. Mouillot, N. Mouquet, and D. Gravel, "The dissimilarity of species interaction networks," *Ecology letters*, vol. 15, no. 12, pp. 1353–1361, 2012.

[14] M. Dale and M.-J. Fortin, "From graphs to spatial graphs," *Annual Review of Ecology, Evolution, and Systematics*, vol. 41, pp. 21–38, 2010.

[15] B. Bollobás, *Modern graph theory*. Springer Science & Business Media, 2013, vol. 184.

[16] N. Wiener, "The theory of prediction," *Modern mathematics for engineers*, 1956.

[17] G. Berkeley, *A treatise concerning the principles of human knowledge*. JB Lippincott & Company, 1881.

[18] J. Aldrich *et al.*, "Correlations genuine and spurious in pearson and yule," *Statistical science*, vol. 10, no. 4, pp. 364–376, 1995.

[19] C. Granger, "Econometrica," vol. 37, p. 424, 1969.

[20] G. Sugihara, R. May, H. Ye, C.-h. Hsieh, E. Deyle, M. Fogarty, and S. Munch, "Detecting causality in complex ecosystems," *Science*, vol. 338, no. 6106, pp. 496–500, 2012. [Online]. Available: https://science.sciencemag.org/content/338/6106/496

[21] M. Ushio, C.-h. Hsieh, R. Masuda, E. R. Deyle, H. Ye, C.-W. Chang, G. Sugihara, and M. Kondoh, "Fluctuating interaction network and time-varying stability of a natural fish community," *Nature*, vol. 554, pp. 360–363, 2018. [Online]. Available: https://doi.org/10.1038/nature25504

[22] H. Ye, E. R. Deyle, G. Luis J., and G. Sugihara, "Distinguishing time-delayed causal interactions using convergent cross mapping," *Scientific Reports*, vol. 5, no. 14750, pp. 1174–1181, 2015. [Online]. Available: https://doi.org/10.1038/srep14750

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

[24] J. T. Lizier and M. Prokopenko, "Differentiating information transfer and causal effect," *The European Physical Journal B*, vol. 73, no. 4, pp. 605–615, 2010.

[25] J. T. Lizier, "Jidt: An information-theoretic toolkit for studying the dynamics of complex systems," *Frontiers in Robotics and AI*, vol. 1, p. 11, 2014.

[26] A. Montalto, L. Faes, and D. Marinazzo, "Mute: A matlab toolbox to compare established and novel estimators of the multivariate transfer entropy," *PLOS ONE*, vol. 9, no. 10, pp. 1–13, 10 2014. [Online]. Available: https://doi.org/10.1371/journal.pone.0109462

[27] F. Abdul Razak and H. J. Jensen, "Quantifying causality in complex systems: Understanding transfer entropy," *PLOS ONE*, vol. 9, no. 6, pp. 1–14, 06 2014. [Online]. Available: https://doi.org/10.1371/journal.pone.0099462

[28] P. Duan, F. Yang, T. Chen, and S. L. Shah, "Direct causality detection via the transfer entropy approach," *IEEE transactions on control systems technology*, vol. 21, no. 6, pp. 2052–2066, 2013.

[29] J. Runge, "Causal network reconstruction from time series: From theoretical assumptions to practical estimation," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 28, no. 7, p. 075310, 2018.

[30] J. L. Servadio and M. Convertino, "Optimal information networks: Application for data-driven integrated health in populations," *Science Advances*, vol. 4, no. 2, 2018.

[31] J. Li and M. Convertino, "Optimal microbiome networks: Macroecology and criticality," *Entropy*, vol. 21, no. 5, p. 506, May 2019. [Online]. Available: http://dx.doi.org/10.3390/e21050506

[32] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999. [Online]. Available: http://science.sciencemag.org/content/286/5439/509

[33] M. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003. [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/S003614450342480

[34] K. Steinhaeuser, N. V. Chawla, and A. R. Ganguly, "An exploration of climate data using complex networks," in *Proceedings of the Third International Workshop on Knowledge Discovery from Sensor Data*, 2009, pp. 23–31.

[35] K. Steinhaeuser, A. R. Ganguly, and N. V. Chawla, "Multivariate and multiscale dependence in the global climate system revealed through complex networks," *Climate dynamics*, vol. 39, no. 3-4, pp. 889–895, 2012.

[36] J. H. Feldhoff, S. Lange, J. Volkholz, J. F. Donges, J. Kurths, and F.-W. Gerstengarbe, "Complex networks for climate model evaluation with application to statistical versus dynamical modeling of south american climate," *Climate dynamics*, vol. 44, no. 5-6, pp. 1567–1581, 2015.

[37] X. Dai, M. Hu, W. Tian, D. Xie, and B. Hu, "Application of epidemiology model on complex networks in propagation dynamics of airspace congestion," *PLOS ONE*, vol. 11, no. 6, pp. 1–11, 06 2016. [Online]. Available: https://doi.org/10.1371/journal.pone.0157945

[38] M. Arquam, A. Singh, and R. Sharma, "Modelling and analysis of delayed sir model on complex network," in *International Conference on Complex Networks and their Applications*. Springer, 2018, pp. 418–430.

[39] R. Albert, "Scale-free networks in cell biology," *Journal of Cell Science*, vol. 118, no. 21, pp. 4947–4957, 2005. [Online]. Available: http://jcs.biologists.org/content/118/21/4947

[40] U. Roy, R. K. Grewal, and S. Roy, "Complex networks and systems biology," in *Systems and Synthetic Biology*. Springer, 2015, pp. 129–150.

[41] P. W. Holland, "Statistics and causal inference," *Journal of the American Statistical Association*, vol. 81, no. 396, pp. 945–960, 1986. [Online].

Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1986.10478354

[42] M. Convertino and L. J. Valverde Jr, "Toward a pluralistic conception of resilience," *Ecological Indicators*, vol. 107, p. 105510, 2019.

[43] A. Batushansky, D. Toubiana, and A. Fait, "Correlation-based network generation, visualization, and analysis as a powerful tool in biological studies: A case study in cancer cell metabolism," *BioMed Research International*, 2016. [Online]. Available: https://doi.org/10.1155/2016/8313272

[44] Z. Ahmed and S. Kumar, "Pearson's correlation coefficient in the theory of networks: A comment," *arXiv e-prints*, p. arXiv:1803.06937, Mar 2018.

[45] D. B. Rubin, "Bayesian inference for causal effects: The role of randomization," *The Annals of Statistics*, vol. 6, no. 1, pp. 34–58, 1978. [Online]. Available: http://www.jstor.org/stable/2958688

[46] S. Y. Kim, S. Imoto, and S. Miyano, "Inferring gene networks from time series microarray data using dynamic Bayesian networks," *Briefings in Bioinformatics*, vol. 4, no. 3, pp. 228–235, 09 2003. [Online]. Available: https://doi.org/10.1093/bib/4.3.228

[47] P. Spirtes, C. Glymour, R. Scheines, S. Kauffman, V. G. Aimale, and F. Wimberly, "Constructing bayesian network models of gene expression networks from microarray data," 2000.

[48] M. Zou and S. D. Conzen, "A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data," *Bioinformatics*, vol. 21, no. 1, pp. 71–79, 08 2004. [Online]. Available: https://doi.org/10.1093/bioinformatics/bth463

[49] Y. Zhang, Z. Deng, H. Jiang, and P. Jia, "Gene regulatory network construction using dynamic bayesian network (dbn) with structure expectation maximization (sem)," in *International Conference on Rough Sets and Knowledge Technology*. Springer, 2006, pp. 402–407.

[50] S. Højsgaard, D. Edwards, and S. Lauritzen, "Gaussian graphical models," in *Graphical Models with R*. Springer, 2012, pp. 77–116.

[51] C. Uhler, "Gaussian Graphical Models: An Algebraic and Geometric Perspective," *arXiv e-prints*, p. arXiv:1707.04345, Jul 2017.

[52] J. B. Grace, *Structural equation modeling and natural systems*. Cambridge University Press, 2006.

[53] P. Barrett, "Structural equation modelling: Adjudging model fit," *Personality and Individual differences*, vol. 42, no. 5, pp. 815–824, 2007.

[54] D. Hooper, J. Coughlan, and M. Mullen, "Structural equation modelling: guidelines for determining model fit. electron j bus res methods 6: 53–60," 2008.

[55] J. B. Grace, D. R. Schoolmaster Jr, G. R. Guntenspergen, A. M. Little, B. R. Mitchell, K. M. Miller, and E. W. Schweiger, "Guidelines for a graph-theoretic implementation of structural equation modeling," *Ecosphere*, vol. 3, no. 8, pp. 1–44, 2012.

[56] H. Yang, W. Yang, J. Zhang, T. Connor, and J. Liu, "Revealing pathways from payments for ecosystem services to socioeconomic outcomes," *Science Advances*, vol. 4, no. 3, 2018. [Online]. Available: https://advances.sciencemag.org/content/4/3/eaao6652

[57] S. D. Mamet, E. Redlick, M. Brabant, E. G. Lamb, B. L. Helgason, K. Stanley, and S. D. Siciliano, "Structural equation modeling of a winnowed soil microbiome identifies how invasive plants re-structure microbial networks," *The ISME Journal*, 2019. [Online]. Available: https://doi.org/10.1038/s41396-019-0407-y

[58] T. Q. Tung, T. Ryu, K. H. Lee, and D. Lee, "Inferring gene regulatory networks from microarray time series data using transfer entropy," in *Twentieth IEEE International Symposium on Computer-Based Medical Systems (CBMS'07)*. IEEE, 2007, pp. 383–388.

[59] A. F. Villaverde, J. Ross, F. Morn, and J. R. Banga, "Mider: Network inference with mutual information distance and entropy reduction," *PLOS ONE*, vol. 9, no. 5, pp. 1–15, 05 2014. [Online]. Available: https://doi.org/10.1371/journal.pone.0096732

[60] J. Sun, D. Taylor, and E. M. Bollt, "Causal network inference by optimal causation entropy," *SIAM Journal on Applied Dynamical Systems*, vol. 14, no. 1, pp. 73–106, 2015.

[61] L. Novelli, P. Wollstadt, P. Mediano, M. Wibral, and J. T. Lizier, "Large-scale directed network inference with multivariate transfer entropy and hierarchical statistical testing," *CoRR*, vol. abs/1902.06828, 2019. [Online]. Available: http://arxiv.org/abs/1902.06828

[62] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. D. Mahecha, J. Muñoz-Marí *et al.*, "Inferring causation from time series in earth system sciences," *Nature communications*, vol. 10, no. 1, pp. 1–13, 2019.

[63] L. Jost, "Entropy and diversity," *Oikos*, vol. 113, no. 2, pp. 363–375, 2006. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2006.0030-1299.14714.x

[64] D. Ruelle and F. Takens, "On the nature of turbulence," *Les rencontres physiciens-mathematiciens de Strasbourg-RCP25*, vol. 12, pp. 1–44, 1971.

[65] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical systems and turbulence, Warwick 1980*. Springer, 1981, pp. 366–381.

[66] A. T. Clark, H. Ye, F. Isbell, E. R. Deyle, J. Cowles, G. D. Tilman, and G. Sugihara, "Spatial convergent cross mapping to detect causal relationships from short time series," *Ecology*, vol. 96, no. 5, pp. 1174–1181, 2015. [Online]. Available: https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/14-1479.1

[67] L. J. A. MacCall, *Can. J. Fish. Aquat. Sci.*, vol. 52, no. 566, 1995.

[68] G. I. Murphy, *Species replacement in marine ecosystems with reference to the California current*. Scripps Institution of Oceanography, 1964.

[69] R. Lasker and A. MacCall, "New ideas on the fluctuations of the clupeoid stocks off california," in *Proceedings of the joint oceanographic assembly 1982General symposia, Ottawa*, 1983.

[70] T. B. et al., *CCOFI Rep*, vol. 33, no. 24, 1992.

[71] K. M. Brander, "Global fish production and climate change," *Proceedings of the National Academy of Sciences*, vol. 104, no. 50, pp. 19 709–19 714, 2007.

[72] C. L. Hein, G. Öhlund, and G. Englund, "Fish introductions reveal the temperature dependence of species interactions," *Proceedings of the Royal Society B: Biological Sciences*, vol. 281, no. 1775, p. 20132641, 2014.

[73] L. Comte and J. D. Olden, "Climatic vulnerability of the worlds freshwater and marine fishes," *Nature Climate Change*, vol. 7, pp. 718–722, 2017. [Online]. Available: https://doi.org/10.1038/nclimate3382

[74] S. Cenci and S. Saavedra, "Uncertainty quantification of the effects of biotic interactions on community dynamics from nonlinear time-series data," *Journal of The Royal Society Interface*, vol. 15, no. 147, p. 20180695, 2018.

[75] M. J. Blaser, Z. G. Cardon, M. K. Cho, J. L. Dangl, T. J. Donohue, J. L. Green, R. Knight, M. E. Maxon, T. R. Northen, K. S. Pollard, and E. L. Brodie, "Toward a predictive understanding of earth's microbiomes to address 21st century challenges," *mBio*, vol. 7, no. 3, 2016. [Online]. Available: https://mbio.asm.org/content/7/3/e00714-16

[76] K. Z. Coyte, J. Schluter, and K. R. Foster, "The ecology of the microbiome: Networks, competition, and stability," *Science*, vol. 350, no. 6261, pp. 663–666, 2015. [Online]. Available: https://science.sciencemag.org/content/350/6261/663

[77] M. van de Guchte, H. M. Blottière, and J. Doré, "Humans as holobionts: implications for prevention and therapy," *Microbiome*, vol. 6, no. 1, p. 81, May 2018. [Online]. Available: https://doi.org/10.1186/s40168-018-0466-8

[78] M. Arumugam, J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls, J.-M. Batto, M. Bertalan, N. Borruel, F. Casellas, L. Fernandez, L. Gautier, T. Hansen, M. Hattori, T. Hayashi, M. Kleerebezem, K. Kurokawa, M. Leclerc, F. Levenez, C. Manichanh, H. B. Nielsen, T. Nielsen, N. Pons, J. Poulain, J. Qin, T. Sicheritz-Ponten, S. Tims, D. Torrents, E. Ugarte, E. G. Zoetendal, J. Wang, F. Guarner, O. Pedersen, W. M. de Vos, S. Brunak, J. Doré, M. Antolín, F. Artiguenave, H. M. Blottiere, M. Almeida, C. Brechot, C. Cara, C. Chervaux, A. Cultrone, C. Delorme, G. Denariaz, R. Dervyn, K. U. Foerstner, C. Friss, M. van de

Guchte, E. Guedon, F. Haimet, W. Huber, J. van Hylckama-Vlieg, A. Jamet, C. Juste, G. Kaci, J. Knol, K. Kristiansen, O. Lakhdari, S. Layec, K. Le Roux, E. Maguin, A. Mérieux, R. Melo Minardi, C. M'rini, J. Muller, R. Oozeer, J. Parkhill, P. Renault, M. Rescigno, N. Sanchez, S. Sunagawa, A. Torrejon, K. Turner, G. Vandemeulebrouck, E. Varela, Y. Winogradsky, G. Zeller, J. Weissenbach, S. D. Ehrlich, P. Bork, and M. C. a. members), "Enterotypes of the human gut microbiome," *Nature*, vol. 473, no. 7346, pp. 174–180, May 2011. [Online]. Available: https://doi.org/10.1038/nature09944

[79] D. Knights, T. L. Ward, C. E. McKinlay, H. Miller, A. Gonzalez, D. McDonald, and R. Knight, "Rethinking enterotypes," *Cell Host and Microbe*, vol. 16, no. 4, pp. 433 – 437, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1931312814003461

[80] R. Caesar, V. Tremaroli, P. Kovatcheva-Datchary, P. D. Cani, and F. Bäckhed, "Crosstalk between gut microbiota and dietary lipids aggravates wat inflammation through tlr signaling," *Cell metabolism*, vol. 22, no. 4, pp. 658–668, Oct 2015, 26321659[pmid]. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/26321659

[81] J. M. Martí, D. Martínez-Martínez, T. Rubio, C. Gracia, M. Peña, A. Latorre, A. Moya, and C. P. Garay, "Health and disease imprinted in the time variability of the human microbiome," *mSystems*, vol. 2, no. 2, 2017. [Online]. Available: https://msystems.asm.org/content/2/2/e00144-16

[82] M. Marsili, I. Mastromatteo, and Y. Roudi, "On sampling and modeling complex systems," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2013, no. 09, 2013. [Online]. Available: http://stacks.iop.org/1742-5468/2013/i=09/a=P09003

[83] M. Layeghifard, D. M. Hwang, and D. S. Guttman, "Disentangling interactions in the microbiome: A network perspective," *Trends in Microbiology*, vol. 25, no. 3, pp. 217 – 228, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0966842X16301858

[84] J. T. Lizier and M. Prokopenko, "Differentiating information transfer and causal effect," *The European Physical Journal B*, vol. 73, no. 4, pp. 605–615, Feb 2010. [Online]. Available: https://doi.org/10.1140/epjb/e2010-00034-5

[85] F. Boschetti, "Models and people: An alternative view of the emergent properties of computational models," *Complexity*, 2015.

[86] T. Zillio, J. Banavar, J. Green, J. Harte, and A. Maritan, "Incipient criticality in ecological communities," *Proceedings of the National Academy of Science*, vol. 105, no. 48, 2008.

[87] M. Convertino, "Neutral metacommunity clustering and sar: River basin vs. 2-d landscape biodiversity patterns," *Ecological Modelling*, vol. 222, no. 11, pp. 1863 – 1879, 2011.

[88] S. Azaele, A. Maritan, E. Bertuzzo, I. Rodriguez-Iturbe, and A. Rinaldo, "Stochastic dynamics of cholera epidemics," *Phys. Rev. E*, vol. 81, no. 051901, 2010.

[89] M. Martinello, J. Hidalgo, A. Maritan, S. di Santo, D. Plenz, and M. A. Muñoz, "Neutral theory and scale-free neural dynamics," *Phys. Rev. X*, vol. 7, p. 041071, 2017. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevX.7.041071

[90] P. Jeraldo, M. Sipos, N. Chia, J. M. Brulc, A. S. Dhillon, M. E. Konkel, C. L. Larson, K. E. Nelson, A. Qu, L. B. Schook, F. Yang, B. A. White, and N. Goldenfeld, "Quantification of the relative roles of niche and neutral processes in structuring gastrointestinal microbiomes," *Proceedings of the National Academy of Sciences*, vol. 109, no. 25, pp. 9692–9698, 2012. [Online]. Available: https://www.pnas.org/content/109/25/9692

[91] R. Levy and E. Borenstein, "Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules," *Proceedings of the National Academy of Sciences*, vol. 110, no. 31, pp. 12 804–12 809, 2013. [Online]. Available: https://www.pnas.org/content/110/31/12804

[92] R. A. Quinn, W. Comstock, T. Zhang, J. T. Morton, R. da Silva, A. Tran, A. Aksenov, L.-F. Nothias, D. Wangpraseurt, A. V. Melnik, G. Ackermann, D. Conrad, I. Klapper, R. Knight, and P. C. Dorrestein, "Niche partitioning of a pathogenic microbiome driven by chemical gradients," *Science Advances*, vol. 4, no. 9, 2018. [Online]. Available: https://advances.sciencemag.org/content/4/9/eaau1908

[93] R. R. Stein, V. Bucci, N. C. Toussaint, C. G. Buffie, G. Rtsch, E. G. Pamer, C. Sander, and J. B. Xavier, "Ecological modeling from time-series inference: Insight into dynamics and stability of intestinal microbiota," *PLOS Computational Biology*, vol. 9, no. 12, pp. 1–11, 12 2013. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1003388

[94] M. Convertino, A. Bockelie, G. A. Kiker, R. Muñoz-Carpena, and I. Linkov, "Shorebird patches as fingerprints of fractal coastline fluctuations due to climate change," *Ecological Processes*, vol. 1, no. 1, 2012. [Online]. Available: https://doi.org/10.1186/2192-1709-1-9

[95] L. Lahti, J. Salojärvi, A. Salonen, M. Scheffer, and W. M. de Vos, "Tipping elements in the human intestinal ecosystem," *Nature Communications*, vol. 5, p. 4344, 2014.

[96] C. L. Gentile and T. L. Weir, "The gut microbiota at the intersection of diet and human health," *Science*, vol. 362, no. 6416, pp. 776–780, 2018. [Online]. Available: https://science.sciencemag.org/content/362/6416/776

[97] D. Gonze, K. Z. Coyte, L. Lahti, and K. Faust, "Microbial communities as dynamical systems," *Current Opinion in Microbiology*, vol. 44, pp. 41 – 49, 2018, microbiota. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1369527418300092

[98] F. Bauchinger, "Self-organized criticality in the gut microbiome," 2015.

[99] J. Hidalgo, J. Grilli, S. Suweis, M. A. Muoz, J. R. Banavar, and A. Maritan, "Information-based fitness and the emergence of criticality in living systems," *Proceedings of the National Academy of Sciences*, vol. 111, pp. 10 095–10 100, 2014.

[100] X. Li and M. Small, "Neuronal avalanches of a self-organized neural network with active-neuron-dominant structure," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 22, no. 2, p. 023104, 2012. [Online]. Available: https://doi.org/10.1063/1.3701946

[101] S.-S. Poil, R. Hardstone, H. D. Mansvelder, and K. Linkenkaer-Hansen, "Critical-state dynamics of avalanches and oscillations jointly emerge from balanced excitation/inhibition in neuronal networks," *Journal of*

Neuroscience, vol. 32, no. 29, pp. 9817–9823, 2012. [Online]. Available: https://www.jneurosci.org/content/32/29/9817

[102] M. Convertino, R. Muneepeerakul, S. Azaele, E. Bertuzzo, A. Rinaldo, and I. Rodriguez-Iturbe, "On neutral metacommunity patterns of river basins at different scales of aggregation," *Water Resources Research*, vol. 45, no. 8, 2009.

[103] S. N. Baldassano and D. S. Bassett, "Topological distortion and reorganized modular structure of gut microbial co-occurrence networks in inflammatory bowel disease," *Scientific Reports*, vol. 6, no. 1, p. 26087, May 2016. [Online]. Available: https://doi.org/10.1038/srep26087

[104] V. Grimm, E. Revilla, U. Berger, F. Jeltsch, W. M. Mooij, S. F. Railsback, H.-H. Thulke, J. Weiner, T. Wiegand, and D. L. DeAngelis, "Pattern-oriented modeling of agent-based complex systems: Lessons from ecology," *Science*, vol. 310, no. 5750, pp. 987–991, 2005. [Online]. Available: https://science.sciencemag.org/content/310/5750/987

[105] K. Faust, F. Bauchinger, B. Laroche, S. de Buyl, L. Lahti, A. D. Washburne, D. Gonze, and S. Widder, "Signatures of ecological processes in microbial community time series," *Microbiome*, vol. 6, no. 1, p. 120, Jun 2018. [Online]. Available: https://doi.org/10.1186/s40168-018-0496-2

[106] A. Hastings, K. C. Abbott, K. Cuddington, T. Francis, G. Gellner, Y.-C. Lai, A. Morozov, S. Petrovskii, K. Scranton, and M. L. Zeeman, "Transient phenomena in ecology," *Science*, vol. 361, no. 6406, 2018. [Online]. Available: https://science.sciencemag.org/content/361/6406/eaat6412

[107] R. A. Chisholm and S. W. Pacala, "Niche and neutral models predict asymptotically equivalent species abundance distributions in high-diversity ecological communities," *Proceedings of the National Academy of Sciences*, vol. 107, no. 36, pp. 15 821–15 825, 2010. [Online]. Available: https://www.pnas.org/content/107/36/15821

[108] A. Durbn, J. J. Abelln, N. Jimnez-Hernndez, A. Artacho, V. Garrigues, V. Ortiz, J. Ponce, A. Latorre, and A. Moya, "Instability of the faecal microbiota in diarrhoea-predominant irritable bowel syndrome," *FEMS*

Microbiology Ecology, vol. 86, no. 3, pp. 581–589, 12 2013. [Online]. Available: https://doi.org/10.1111/1574-6941.12184

[109] M. Convertino, F. Simini, F. Catani, I. Linkov, and G. Kiker, "Power-law of aggregate-size spectra in natural systems," *EAI Endorsed Trans. Complex Syst.*, vol. 1, p. e2, 2013.

[110] C. James, S. Azaele, A. Maritan, and F. Simini, "Zipf's and taylor's laws," *Phys. Rev. E*, vol. 98, p. 032408, Sep 2018. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.98.032408

[111] J. T. Lizier, "Jidt: An information-theoretic toolkit for studying the dynamics of complex systems," *Frontiers in Robotics and AI*, vol. 1, p. 11, 2014. [Online]. Available: https://www.frontiersin.org/article/10.3389/frobt.2014.00011

[112] R. G. James, N. Barnett, and J. P. Crutchfield, "Information flows? a critique of transfer entropies," *Phys. Rev. Lett.*, vol. 116, p. 238701, Jun 2016. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.116.238701

[113] C. Borile, M. A. Muñoz, S. Azaele, J. R. Banavar, and A. Maritan, "Spontaneously broken neutral symmetry in an ecological system," *Phys. Rev. Lett.*, vol. 109, p. 038102, Jul 2012. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.109.038102

[114] R. Hanel, S. Thurner, and M. Gell-Mann, "How multiplicity determines entropy and the derivation of the maximum entropy principle for complex systems," *Proceedings of the National Academy of Sciences*, vol. 111, no. 19, pp. 6905–6910, 2014. [Online]. Available: https://www.pnas.org/content/111/19/6905

[115] D. W. Rivett and T. Bell, "Abundance determines the functional role of bacterial phylotypes in complex communities," *Nature Microbiology*, vol. 3, no. 7, pp. 767–772, Jul 2018. [Online]. Available: https://doi.org/10.1038/s41564-018-0180-0

[116] J. R. Banavar, A. Maritan, and A. Rinaldo, "Size and form in efficient transportation networks," *Nature*, vol. 399, no. 6732, pp. 130–132, May 1999. [Online]. Available: https://doi.org/10.1038/20144

[117] D. Vandeputte, G. Kathagen, K. D'hoe, S. Vieira-Silva, M. Valles-Colomer, J. Sabino, J. Wang, R. Y. Tito, L. De Commer, Y. Darzi, S. Vermeire, G. Falony, and J. Raes, "Quantitative microbiome profiling links gut community variation to microbial load," *Nature*, vol. 551, no. 7681, pp. 507–511, Nov 2017. [Online]. Available: https://doi.org/10.1038/nature24460

[118] C. Mellin, C. J. A. Bradshaw, D. A. Fordham, and M. J. Caley, "Strong but opposing ¡i¿β¡/i¿-diversity–stability relationships in coral reef fish communities," *Proceedings of the Royal Society B: Biological Sciences*, vol. 281, no. 1777, p. 20131993, 2014. [Online]. Available: https://royalsocietypublishing.org/doi/abs/10.1098/rspb.2013.1993

[119] J. R. Zaneveld, D. E. Burkepile, A. A. Shantz, C. E. Pritchard, R. McMinds, J. P. Payet, R. Welsh, A. M. S. Correa, N. P. Lemoine, S. Rosales, C. Fuchs, J. A. Maynard, and R. V. Thurber, "Overfishing and nutrient pollution interact with temperature to disrupt coral reefs down to microbial scales," *Nature Communications*, vol. 7, p. 11833, 2016.

[120] T. J. Matthews and R. J. Whittaker, "Review: On the species abundance distribution in applied ecology and biodiversity management," *Journal of Applied Ecology*, vol. 52, no. 2, pp. 443–454, 2015. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/1365-2664.12380

[121] G. Csányi and B. Szendrői, "Fractal–small-world dichotomy in real-world networks," *Phys. Rev. E*, vol. 70, p. 016122, Jul 2004. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.70.016122

[122] J. P. DeLong, J. G. Okie, M. E. Moses, R. M. Sibly, and J. H. Brown, "Shifts in metabolic scaling, production, and efficiency across major evolutionary transitions of life," *Proceedings of the National Academy of Sciences*, vol. 107, no. 29, pp. 12 941–12 945, 2010. [Online]. Available: https://www.pnas.org/content/107/29/12941

[123] K. R. Foster, J. Schluter, K. Z. Coyte, and S. Rakoff-Nahoum, "The evolution of the host microbiome as an ecosystem on a leash," *Nature*, vol. 548, no. 7665, pp. 43–51, Aug 2017. [Online]. Available: https://doi.org/10.1038/nature23292

[124] M. E. Newman, "Power laws, pareto distributions and zipf's law," *Contemporary Physics*, vol. 46, no. 5, 2005.

[125] R. Albert, H. Jeong, and A.-L. Barabasi, "Error and attack tolerance of complex networks," *Nature*, vol. 406, no. 6794, pp. 378–382, 2000.

[126] F. Isbell, V. Calcagno, A. Hector, J. Connolly, W. S. Harpole, P. B. Reich, M. Scherer-Lorenzen, B. Schmid, D. Tilman, J. van Ruijven, A. Weigelt, B. J. Wilsey, E. S. Zavaleta, and M. Loreau, "High plant diversity is needed to maintain ecosystem services," *Nature*, vol. 477, no. 7363, pp. 199–202, Sep 2011. [Online]. Available: https://doi.org/10.1038/nature10282

[127] A. S. Mori, F. Isbell, and R. Seidl, "-diversity, community assembly, and ecosystem functioning," *Trends in Ecology and Evolution*, vol. 33, no. 7, pp. 549 – 564, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169534718300909

[128] M. Convertino, R. Muñoz-Carpena, G. A. Kiker, and S. G. Perz, "Design of optimal ecosystem monitoring networks: hotspot detection and biodiversity patterns," *Stochastic Environmental Research and Risk Assessment*, vol. 29, no. 4, pp. 1085–1101, May 2015. [Online]. Available: https://doi.org/10.1007/s00477-014-0999-8

[129] F. Tria, V. Loreto, and V. Servedio, "Zipfs, heaps and taylors laws are determined by the expansion into the adjacent possible," *Entropy*, vol. 20, no. 10, p. 752, Sep 2018. [Online]. Available: http://dx.doi.org/10.3390/e20100752

[130] G. B. West, J. H. Brown, and B. J. Enquist, "A general model for the origin of allometric scaling laws in biology," *Science*, vol. 276, no. 5309, pp. 122–126, 1997. [Online]. Available: https://science.sciencemag.org/content/276/5309/122

[131] L. F. Seoane and R. Solé, "Systems poised to criticality through pareto selective forces," *arXiv: Statistical Mechanics*, 2015.

[132] A. Tendler, A. Mayo, and U. Alon, "Evolutionary tradeoffs, pareto optimality and the morphology of ammonite shells," *BMC Systems Biology*, vol. 9, no. 1, p. 12, Mar 2015. [Online]. Available: https://doi.org/10.1186/s12918-015-0149-z

[133] L. Koçillari, P. Fariselli, A. Trovato, F. Seno, and A. Maritan, "Signature of pareto optimization in the escherichia coli proteome," *Scientific Reports*, vol. 8, no. 1, p. 9141, Jun 2018. [Online]. Available: https://doi.org/10.1038/s41598-018-27287-3

[134] S. Suweis, J. Grilli, J. R. Banavar, S. Allesina, and A. Maritan, "Effect of localization on the stability of mutualistic ecological networks," *Nature Communications*, vol. 6, no. 1, p. 10179, Dec 2015. [Online]. Available: https://doi.org/10.1038/ncomms10179

[135] A. M. Kilpatrick and A. R. Ives, "Species interactions can explain taylor's power law for ecological time series," *Nature*, vol. 422, no. 6927, pp. 65–68, Mar 2003. [Online]. Available: https://doi.org/10.1038/nature01471

[136] J. Grilli, G. Barabás, M. J. Michalska-Smith, and S. Allesina, "Higher-order interactions stabilize dynamics in competitive network models," *Nature*, vol. 548, no. 7666, pp. 210–213, Aug 2017. [Online]. Available: https://doi.org/10.1038/nature23273

[137] J. M. Levine, J. Bascompte, P. B. Adler, and S. Allesina, "Beyond pairwise mechanisms of species coexistence in complex communities," *Nature*, vol. 546, no. 7656, pp. 56–64, Jun 2017. [Online]. Available: https://doi.org/10.1038/nature22898

[138] R. Quax, A. Apolloni, and P. M. A. Sloot, "The diminishing role of hubs in dynamical processes on complex networks," *Journal of The Royal Society Interface*, vol. 10, no. 88, p. 20130568, 2013. [Online]. Available: https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2013.0568

[139] M. Granovetter, "The strength of weak ties: A network theory revisited," *Sociological Theory*, vol. 1, pp. 201–233, 2020/10/21/ 1983, full publication date: 1983. [Online]. Available: https://doi.org/10.2307/202051

[140] B. Garca-Jimnez, T. de la Rosa, and M. D. Wilkinson, "MDPbiome: microbiome engineering through prescriptive perturbations," *Bioinformatics*, vol. 34, no. 17, pp. i838–i847, 09 2018. [Online]. Available: https://doi.org/10.1093/bioinformatics/bty562

[141] H. Toju, K. G. Peay, M. Yamamichi, K. Narisawa, K. Hiruma, K. Naito, S. Fukuda, M. Ushio, S. Nakaoka, Y. Onoda, K. Yoshida, K. Schlaeppi, Y. Bai, R. Sugiura, Y. Ichihashi, K. Minamisawa, and E. T. Kiers, "Core microbiomes for sustainable agroecosystems," *Nature Plants*, vol. 4, no. 5, pp. 247–257, May 2018. [Online]. Available: https://doi.org/10.1038/s41477-018-0139-4

[142] A. P. Allen, T. G. Dinan, G. Clarke, and J. F. Cryan, "A psychology of the human braingutmicrobiome axis," *Social and Personality Psychology Compass*, vol. 11, no. 4, p. e12309, 2017, e12309 SPCO-0798.R1. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/spc3.12309

[143] S. Wang and M. Loreau, "Biodiversity and ecosystem stability across scales in metacommunities," *Ecology Letters*, vol. 19, no. 5, pp. 510–518, 2016. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/ele.12582

[144] M. Scheffer, S. R. Carpenter, T. M. Lenton, J. Bascompte, W. Brock, V. Dakos, J. van de Koppel, I. A. van de Leemput, S. A. Levin, E. H. van Nes, M. Pascual, and J. Vandermeer, "Anticipating critical transitions," *Science*, vol. 338, no. 6105, pp. 344–348, 2012. [Online]. Available: https://science.sciencemag.org/content/338/6105/344

[145] U. FAO, "The state of world fisheries and aquaculture 2016. contributing to food security and nutrition for all," 2016.

[146] FAO, "The state of world fisheries and aquaculture 2020. sustainability in action," *Food and Agriculture Organization of the United Nations*, 2020.

[147] H. Ritchie, "Seafood production," *Our World in Data*, 2019, https://ourworldindata.org/seafood-production.

[148] L. Cheng, J. Abraham, Z. Hausfather, and K. E. Trenberth, "How fast are the oceans warming?" *Science*, vol. 363, no. 6423, pp. 128–129, 2019. [Online]. Available: https://science.sciencemag.org/content/363/6423/128

[149] C. M. Free, J. T. Thorson, M. L. Pinsky, K. L. Oken, J. Wiedenmann, and O. P. Jensen, "Impacts of historical warming on marine fisheries production," *Science*, vol. 363, no. 6430, pp. 979–983, 2019. [Online]. Available: https://science.sciencemag.org/content/363/6430/979

[150] W. W. L. Cheung, R. Watson, and D. Pauly, "Signature of ocean warming in global fisheries catch," *Nature*, vol. 497, no. 7449, pp. 365–368, May 2013. [Online]. Available: https://doi.org/10.1038/nature12156

[151] A. Hunter, D. C. Speirs, and M. R. Heath, "Population density and temperature correlate with long-term trends in somatic growth rates and maturation schedules of herring and sprat," *PLOS ONE*, vol. 14, no. 3, pp. 1–22, 03 2019. [Online]. Available: https://doi.org/10.1371/journal.pone.0212176

[152] H. du Pontavice, D. Gascuel, G. Reygondeau, A. Maureaud, and W. W. L. Cheung, "Climate change undermines the global functioning of marine food webs," *Global Change Biology*, vol. 26, no. 3, pp. 1306–1318, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.14944

[153] C.-Z. Yao and H.-Y. Li, "Effective transfer entropy approach to information flow among epu, investor sentiment and stock market," *Frontiers in Physics*, vol. 8, p. 206, 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/fphy.2020.00206

[154] C. Mora, D. P. Tittensor, S. Adl, A. G. B. Simpson, and B. Worm, "How many species are there on earth and in the ocean?" *PLOS Biology*, vol. 9, no. 8, pp. 1–8, 08 2011. [Online]. Available: https://doi.org/10.1371/journal.pbio.1001127

[155] P. García-Palacios, N. Gross, J. Gaitán, and F. T. Maestre, "Climate mediates the biodiversity–ecosystem stability relationship globally," *Proceedings of the National Academy of Sciences*, vol. 115, no. 33, pp. 8400–8405, 2018. [Online]. Available: https://www.pnas.org/content/115/33/8400

[156] S. L. Pimm, "The complexity and stability of ecosystems," *Nature*, vol. 307, no. 5949, pp. 321–326, Jan 1984. [Online]. Available: https://doi.org/10.1038/307321a0

[157] V. Grimm and C. Wissel, "Babel, or the ecological stability discussions: an inventory and analysis of terminology and a guide for avoiding confusion," *Oecologia*, vol. 109, no. 3, pp. 323–334, Feb 1997. [Online]. Available: https://doi.org/10.1007/s004420050090

[158] M. Loreau and C. de Mazancourt, "Biodiversity and ecosystem stability: a synthesis of underlying mechanisms," *Ecology Letters*, vol. 16, no. s1, pp. 106–115, 2013. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/ele.12073

[159] A. R. Ives and S. R. Carpenter, "Stability and diversity of ecosystems," *Science*, vol. 317, no. 5834, pp. 58–62, 2007. [Online]. Available: https://science.sciencemag.org/content/317/5834/58

[160] R. Masuda, M. Shiba, Y. Yamashita, M. Ueno, Y. Kai, A. Nakanishi, M. Torikoshi, and M. Tanaka, "Fish assemblages associated with three types of artificial reefs: density of assemblages and possible impacts on adjacent fish abundance," *Fishery Bulletin*, vol. 108, no. 2, pp. 162–173, 2010. [Online]. Available: http://aquaticcommons.org/id/eprint/8755

[161] K. S. Suzuki, K. W. Suzuki, E. Kumakura, K. Sato, Y. Oe, T. Sato, H. Sawada, R. Masuda, and Y. Nogata, "Seasonal alternation of the ontogenetic development of the moon jellyfish aurelia coerulea in maizuru bay, japan," *PLOS ONE*, vol. 14, no. 11, pp. 1–22, 11 2019. [Online]. Available: https://doi.org/10.1371/journal.pone.0225513

[162] C.-W. Chang, M. Ushio, and C.-h. Hsieh, "Empirical dynamic modeling for beginners," *Ecological Research*, vol. 32, no. 6, pp. 785–796, Nov 2017. [Online]. Available: https://doi.org/10.1007/s11284-017-1469-9

[163] C. Gillespie, "Fitting heavy tailed distributions: The powerlaw package," *Journal of Statistical Software, Articles*, vol. 64, no. 2, pp. 1–16, 2015. [Online]. Available: https://www.jstatsoft.org/v064/i02

[164] G. Jóhannesson, G. Bjrnsson, and E. H. Gudmundsson, "Afterglow light curves and broken power laws: A statistical study," *The Astrophysical Journal*, vol. 640, no. 1, pp. L5–L8, feb 2006. [Online]. Available: https://doi.org/10.1086%2F503294

[165] G. A. Begg and J. R. Waldman, "An holistic approach to fish stock identification," *Fisheries research*, vol. 43, no. 1-3, pp. 35–44, 1999.

[166] E. L. Cadima, *Fish stock assessment manual*. Food & Agriculture Org., 2003, no. 393.

[167] F. A. La Sorte, D. Fink, W. M. Hochachka, J. P. DeLong, and S. Kelling, "Spring phenology of ecological productivity contributes to the use of looped migration strategies by birds," *Proceedings of the Royal Society B: Biological Sciences*, vol. 281, no. 1793, p. 20140984, 2014. [Online]. Available: https://royalsocietypublishing.org/doi/abs/10.1098/rspb.2014.0984

[168] V. N. Gudivada, D. Rao, and V. V. Raghavan, "Big data driven natural language processing research and applications," in *Handbook of Statistics*. Elsevier, 2015, vol. 33, pp. 203–238.

[169] K. Hlaváčková-Schindler, M. Paluš, M. Vejmelka, and J. Bhattacharya, "Causality detection based on information-theoretic approaches in time series analysis," *Physics Reports*, vol. 441, no. 1, pp. 1–46, 2007.

[170] R. Vicente, M. Wibral, M. Lindner, and G. Pipa, "Transfer entropy—a model-free measure of effective connectivity for the neurosciences," *Journal of Computational Neuroscience*, vol. 30, no. 1, pp. 45–67, Feb 2011. [Online]. Available: https://doi.org/10.1007/s10827-010-0262-3

[171] L. Stone, "The feasibility and stability of large complex biological networks: a random matrix approach," *Scientific Reports*, vol. 8, no. 1, p. 8246, 2018. [Online]. Available: https://doi.org/10.1038/s41598-018-26486-2

[172] D. Grady, C. Thiemann, and D. Brockmann, "Robust classification of salient links in complex networks," *Nature Communications*, vol. 3, no. 1, p. 864, May 2012. [Online]. Available: https://doi.org/10.1038/ncomms1847

[173] E. S. Poloczanska, C. J. Brown, W. J. Sydeman, W. Kiessling, D. S. Schoeman, P. J. Moore, K. Brander, J. F. Bruno, L. B. Buckley, M. T. Burrows *et al.*, "Global imprint of climate change on marine life," *Nature Climate Change*, vol. 3, no. 10, pp. 919–925, 2013.

[174] C. C. C. Wabnitz, V. W. Y. Lam, G. Reygondeau, L. C. L. Teh, D. Al-Abdulrazzak, M. Khalfallah, D. Pauly, M. L. D. Palomares, D. Zeller, and W. W. L. Cheung, "Climate change impacts on marine biodiversity, fisheries and society in the arabian gulf," *PLOS ONE*, vol. 13, no. 5, pp. 1–26, 05 2018. [Online]. Available: https://doi.org/10.1371/journal.pone.0194537

[175] M. Bastian, S. Heymann, M. Jacomy *et al.*, "Gephi: an open source software for exploring and manipulating networks." *Icwsm*, vol. 8, no. 2009, pp. 361–362, 2009.

[176] V. Pareto, *Manual of political economy: a critical and variorum edition*. OUP Oxford, 2014.

[177] L. Nunney, "The stability of complex model ecosystems," *The American Naturalist*, vol. 115, no. 5, pp. 639–649, 1980, full publication date: May, 1980. [Online]. Available: http://www.jstor.org/stable/2460683

[178] X. CHEN and J. E. COHEN, "Global stability, local stability and permanence in model food webs," *Journal of Theoretical Biology*, vol. 212, no. 2, pp. 223 – 235, 2001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022519301923707

[179] B. S. Goh, "Global stability in many-species systems," *The American Naturalist*, vol. 111, no. 977, pp. 135–143, 1977, full publication date: Jan. - Feb., 1977. [Online]. Available: http://www.jstor.org/stable/2459985

[180] Ö. Ak Gümüs, "Global and local stability analysis in a nonlinear discrete-time population model," *Advances in Difference Equations*, vol. 2014, no. 1, p. 299, 2014. [Online]. Available: https://doi.org/10.1186/1687-1847-2014-299

[181] P. J. Woolf, *Chemical Process Dynamics and Controls*. openmichigan, 2009.

[182] K. Mabuchi, M. Miya, Y. Azuma, and M. Nishida, "Independent evolution of the specialized pharyngeal jaw apparatus in cichlid and labrid fishes," *BMC Evolutionary Biology*, vol. 7, no. 1, p. 10, Jan 2007. [Online]. Available: https://doi.org/10.1186/1471-2148-7-10

[183] O. Dangles, C. Carpio, A. R. Barragan, J.-L. Zeddam, and J.-F. Silvain, "Temperature as a key driver of ecological sorting among invasive pest species in the tropical andes," *Ecological Applications*, vol. 18, no. 7, pp. 1795–1809, 2008. [Online]. Available: https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/07-1638.1

[184] K. Kitayama, M. Ushio, and S.-I. Aiba, "Temperature is a dominant driver of distinct annual seasonality of leaf litter production of equatorial

tropical rain forests," *Journal of Ecology*, vol. n/a, no. n/a, 2020. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/1365-2745.13500

[185] A. S. Brierley and M. J. Kingsford, "Impacts of climate change on marine organisms and ecosystems," *Current biology*, vol. 19, no. 14, pp. R602–R614, 2009.

Appendix A

Supplement for Chapter 2

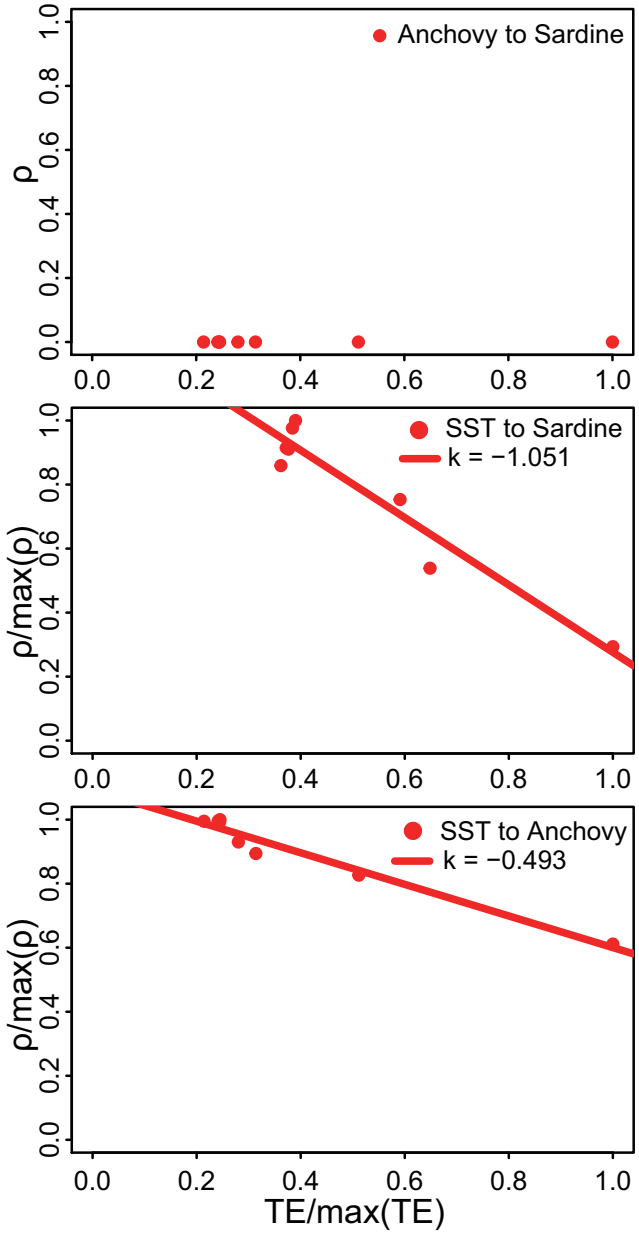

Figure A.1: **Relationship between suboptimal ρ and TE**. Suboptimal ρ-s and TEs for all selected library length L of time series and the time delay u between variables (sardine-anchovy, sardine and SST, anchovy and SST from top to bottom), respectively, are shown.

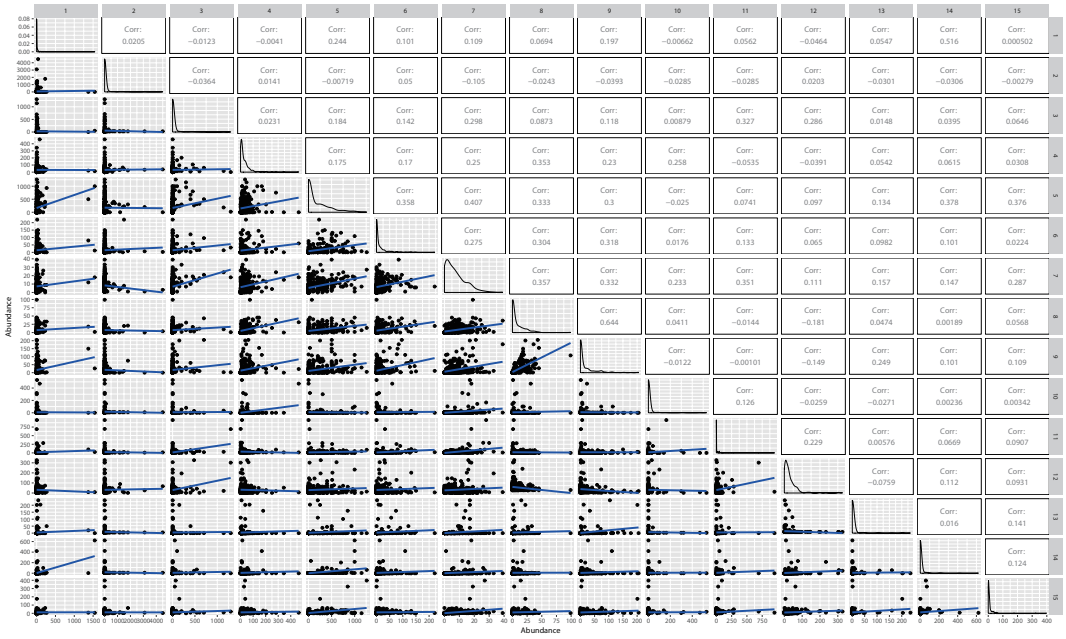

Figure A.2: **Interspecies abundance pattern**. Abundance-abundance patterns of all species in the Maizuru bay independent of time. The higher the correlation coefficient the higher the divergence and asynchronicity between abundance time series, and the higher TE (see Figure 2.7 for species from 4 to 9). ”Mirage” correlation between abundance of species (without considering the delay between autocorrelated values) implies non-linearity and potentially strong causality/physical interaction as demonstrated in Figure 3.4 by the mathematical model results. Vice-versa, lack of correlation or low correlation implies linearity into the dynamics and potentially low causality/physical interaction. TE is advantageous because it is asymmetrical while interspecies correlation is symmetrical, yet not allowing one to capture the directional interaction between species.

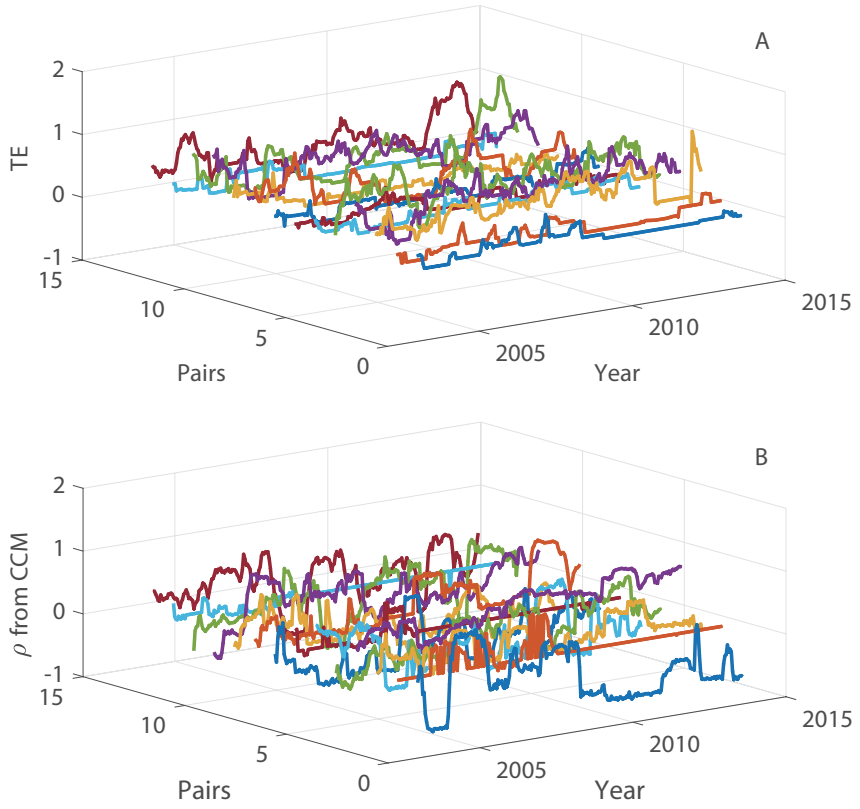

Figure A.3: **Time-varying interspecies interactions via OIF and CCM model for the Maizuru bay fish community.** Interspecies interactions for the 14 pairs of fish species listed in Table. A.2 quantified via (A) OIF and (B) CCM models. The y-axis indicates the time period from 2002 to 2014 over which the abundance of species was sampled every two weeks.

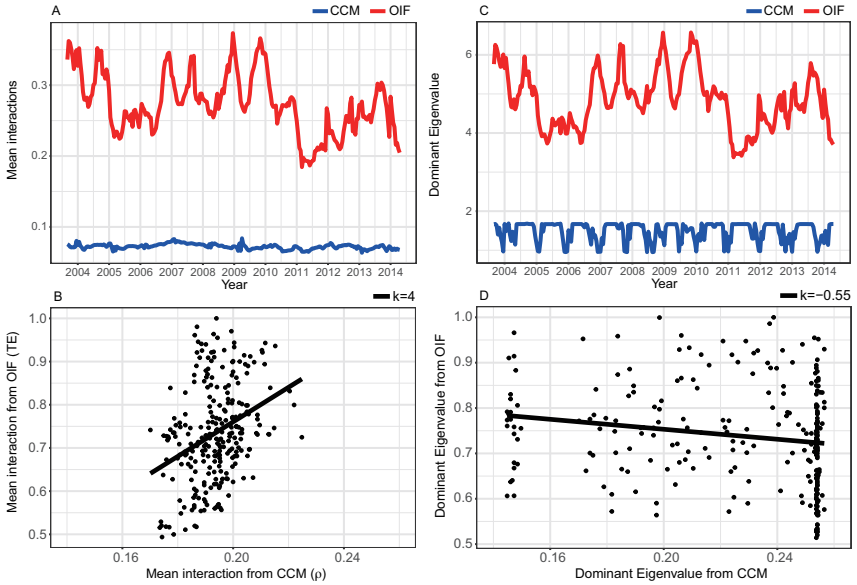

Figure A.4: **Mean interaction strength and dominant eigenvalue from OIF and CCM interaction matrix.** (A) Average of all species-species interactions over time; (OIF and CCM model estimates in red and blue). (B) Atemporal relationship between mean interactions from OIF and CCM models. (C) Dominant eigenvalue corresponding to the highest frequency in interactions fluctuations (OIF and CCM model estimates in red and blue); the real part of the dominant eigenvalue of the interaction matrix at each time point is reported and represents a potential dynamical stability. (D) Atemporal relationship between dominant eigenvalue from OIF and CCM models. For (B) and (D) the slope k is the liner regression coefficient.

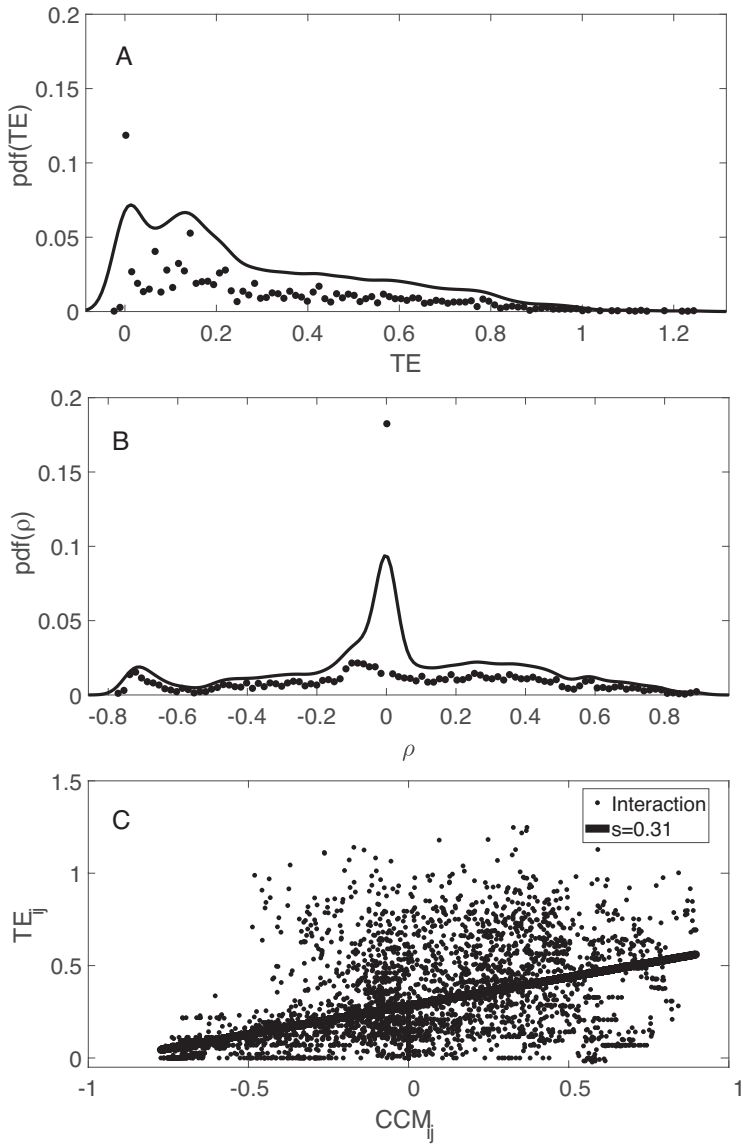

Figure A.5: **Distribution of pairwise species interactions from CCM and OIF**. (A) pdf of TE for all species interactions in the period 2002-2014. (B) pdf of ρ for all species interactions in the period 2002-2014. (C) TE and CCM calculated for each species pair ij considering each time period where abundance time series are updated every two weeks in the period 2002-2014. The number of estimated TEs and ρ (i.e., 3584) is given by the number of observations N (i.e., 256) multiplied by the number of species pairs (i.e., 14).

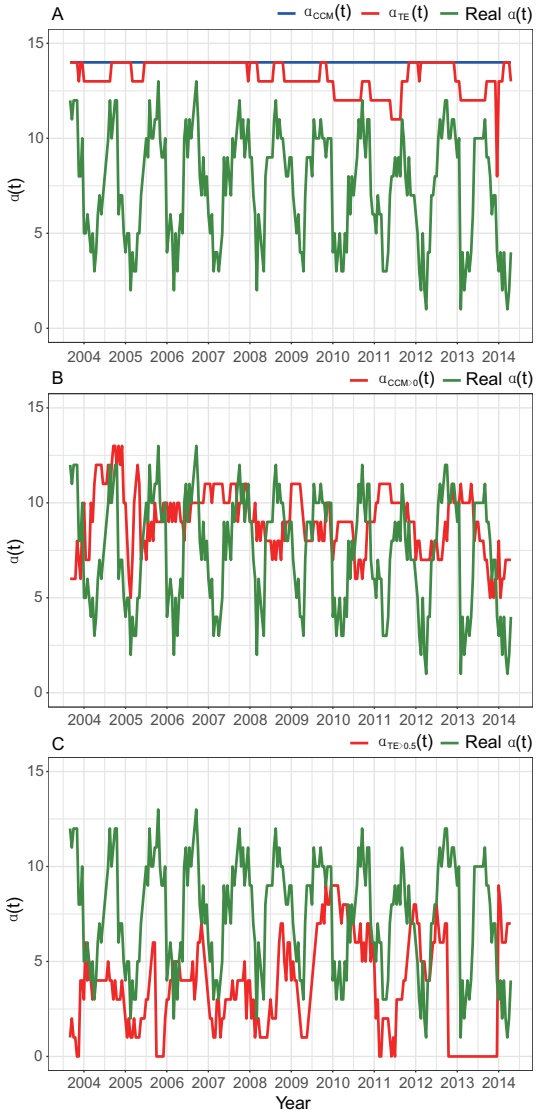

Figure A.6: **Predicted α-diversity via CCM and OIF versus taxonomic diversity.** (A) "Real" taxonomic α-diversity (green line), and inferred temporal α-diversity from CCM and OIF (red lines) without setting any threshold on the magnitude of species interactions (ρ and TE). (B) Inferred α-diversity after setting the threshold to zero for CCM ρ (see Figure A.5 for pdf of ρ where ρ can be negative). (C) Inferred α-diversity from OIF after setting the threshold to 0.5 for TE (see Figure A.5 for pdf of TE). TEs are obtained from JIDT using a time delay $u = 1$ that corresponds to 2 weeks. Sample data are provided every two weeks for 12 years (see Figure 2.7).

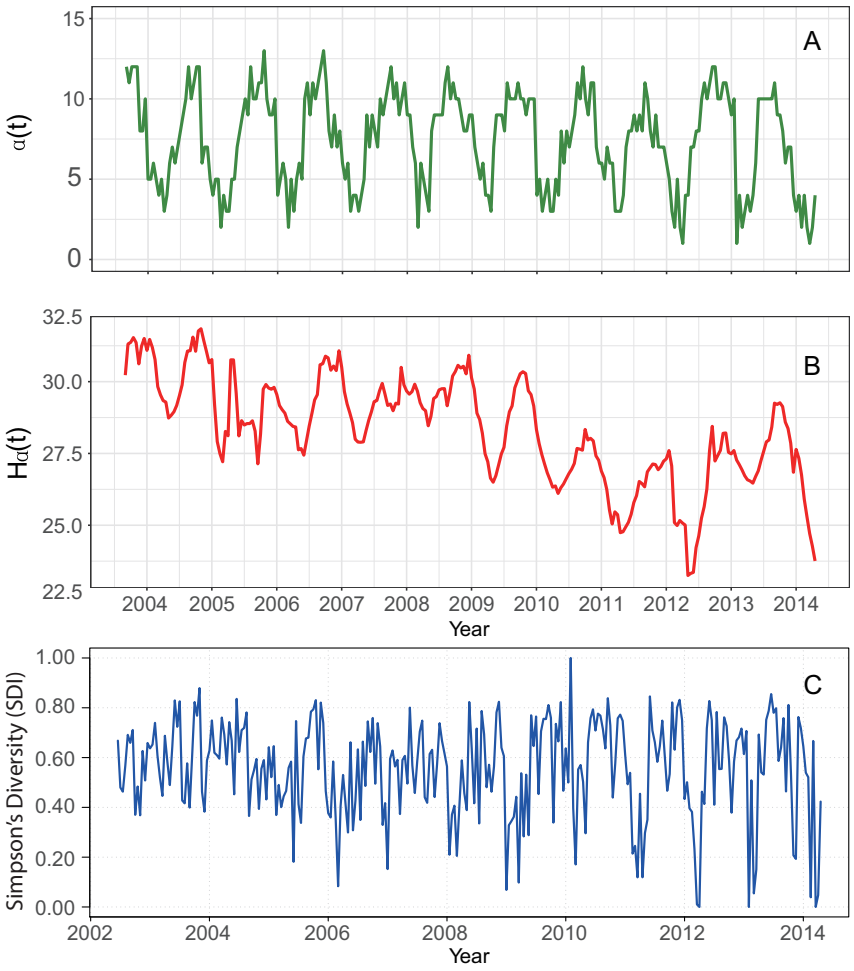

Figure A.7: **Biodiversity indicators over time for the Maizuru bay fish community.** (A) Taxonomic α-diversity as count of diverse species. (B) Shannon diversity index $H_\alpha(t)$ based on the pdf of population abundance of species at each time step (Eq. 2.8). (C) Simpson's diversity index (SDI) over time that measures diversity difference based on population abundance at adjacent time steps (Eq. 2.7). $H_\alpha(t)$ is capturing more the trend of biodiversity that is in this case decaying over time in terms of abundance but increasing in regularity because of the lower entropy. SDI is more related to fluctuations whose periodicity is getting more stable in this ecosystem and observable in the autocorrelation of α.

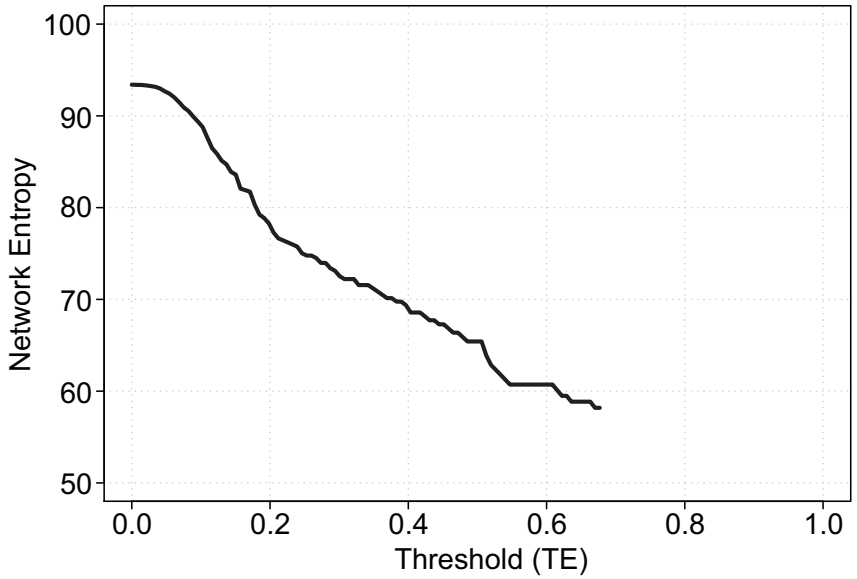

Figure A.8: **Network entropy dependent on the TE threshold.** Network entropy dependent on the pairwise information flow (TE) between species. Network entropy is defined as the sum of Shannon entropies of all species (considering abundance) and TEs of all pairwise species interactions.

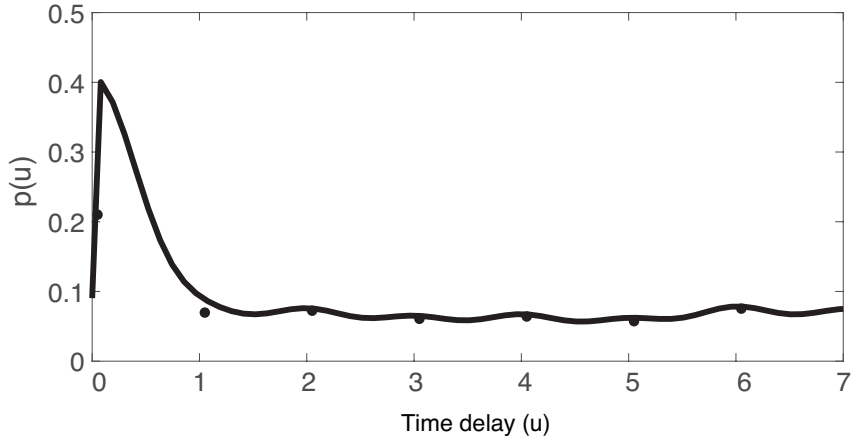

Figure A.9: **Pdf of time delay for TE**. u is the time delay that minimizes the statistical distance defined as $\exp^{-MI(X,Y)}$, where MI is the Mutual Information between species X and Y. The elementary unit or resolution of time delay $u = 1$ corresponds to the species sampling of two weeks.

Species	Fish Stock	Native/Invasive	Conservation Status
1. *Aurelia aurita* (Moon jellyfish)	Yes	Native	NE
2. *Engraulis japonicus* (Japanese anchovy)	Yes	Native	LC
3. *Plotosus lineatus japonicus* (Sea catfish)	No	Invasive	NE
4. *Sebastes inermis* (Black snapper)	Yes	Native	LC
5. *Trachurus japonicus* (Horse mackerel)	Yes	Native	NT
6. *Girella punctata* (Blackeye seabream)	Yes	Native	NE
7. *Pseudolabrus sieboldi* (Wrasse)	Yes	Native	LC
8. *Halichoeres poecilopterus* (Rainbow wrasse)	Yes	Native	LC
9. *Halichoeres tenuispinnis* (Chinese wrasse)	No	Invasive	LC
10. *Chaenogobius gulosus* (Goby)	No	Native	NE
11. *Pterogobius zonoleucus* (Blue/Yellow striped Goby)	No	Native	LC
12. *Tridentiger trigonocephalus* (Chameleon goby)	No	Native	NE
13. *Siganus fuscescens* (Rabbitfish)	Yes	Invasive	LC
14. *Sphyraena pinguis* (Red barracuda)	Yes	Native	NE
15. *Rudarius ercodes* (Pigmy filefish)	No	Invasive	LC

Table A.1: Species ID considering the Maizuru dataset, scientific and common name, categorization in terms of fish stock, location endemicity (native/invasive) and reported IUCN conservation status (up to December 2020).

Number	**Source**	**Target**	$\langle\rho\rangle$	$\langle TE\rangle$
1	S. pinguis	T. japonicus	-0.2451	0.0886
2	T. japonicus	S. pinguis	0.0701	0.1029
3	T. japonicus	Aurelia a.	0.1416	0.4881
4	H. tenuispinis	P. poecilepterus	0.5860	0.5781
5	P. poecilepterus	S. cheni	0.0718	0.6516
6	P. l. japonicus	P. sieboldi	-0.1407	0.1387
7	T. trigonocephalus	C. gulosus	0.0241	0.1393
8	S. fuscescens	P. poecilepterus	-0.3727	0.1469
9	G. punctata	P. zonoleucus	0.1606	0.3199
10	P. l. japonicus	T. trigonocephalus	-0.2120	0.1253
11	R. ercodes	T. japonicus	0.1606	0.4739
12	P. zonoleucus	R. ercodes	0.0210	0.3427
13	P. zonoleucus	C. gulosus	-0.0124	0.0990
14	P. zonoleucus	P. sieboldi	0.0828	0.3734

Table A.2: ρ and transfer entropy of 14 pairs of fish species.

Species	Entropy	OTE	ITE	Mean	Std
1. Aurelia aurita	2.5031	6.4767	9.612	23.826	126.05
2. Engraulis japonicus	0.5068	8.1583	3.4124	68.056	379.3
3. Plotosus lineatus japonicus	0.5472	7.0584	3.1744	28.253	133.03
4. Sebastes inermis	0.3261	6.16	10.524	31.874	50.512
5. Trachurus japonicus	0.9996	6.3931	8.029	174.94	258.7
6. Girella punctata	0.9914	6.0511	7.7665	14.863	29.911
7. Pseudolabrus sieboldi	0.5570	6.2727	10.308	7.3333	6.9701
8. Halichoeres poecilopterus	0.9803	6.1276	6.978	7.5649	11.755
9. Halichoeres tenuispinnis	0.9903	6.6597	5.7889	17.575	33.523
10. Chaenogobius gulosus	0.4852	8.0477	2.5391	8.9386	51.713
11. Pterogobius zonoleucus	0.9379	6.2296	6.8386	18.542	77.554
12. Tridentiger trigonocephalus	0.2022	6.9856	11.227	31.458	43.289
13. Siganus fuscescens	0.4855	7.0709	3.7683	4.6456	25.866
14. Sphyraena pinguis	0.3922	8.1649	2.3686	8.7614	50.779
15. Rudarius ercodes	0.5852	6.0053	9.527	12.142	33.463

Table A.3: Shannon entropy (Entropy), outgoing transfer entropy (OTE), incoming transfer entropy (ITE), mean relative abundance (Mean) and standard deviation (Std).

Appendix B

Supplement for Chapter 3

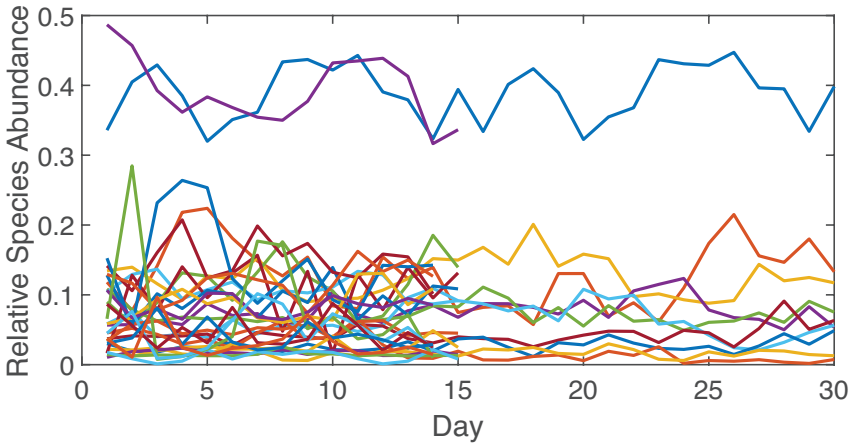

Figure B.1: **RSA time series for all species.**. The RSA of species is reported over time independently of the microbiome state.

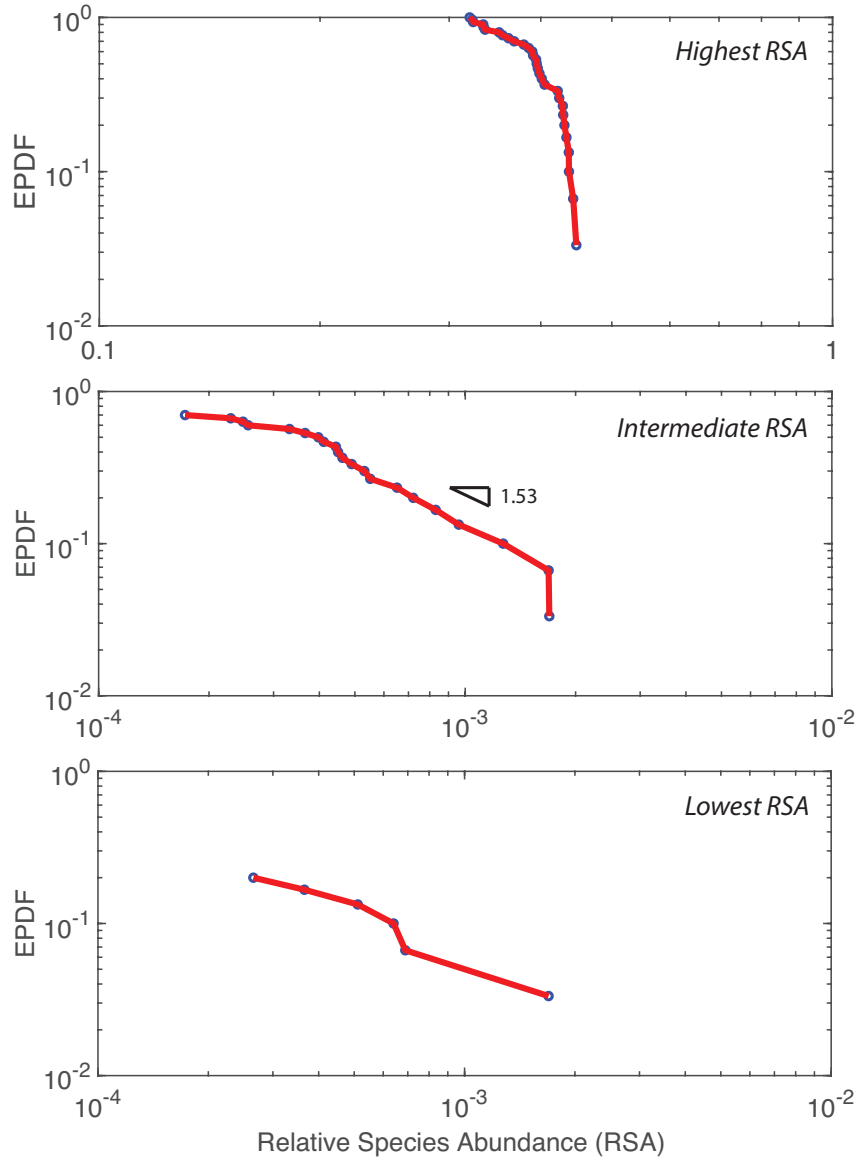

Figure B.2: **Exceedance probability of RSA for all species.** The epdf of RSA is plotted for the top 10 highest RSA, intermediate 10 RSA, and the least 10 RSA species. A power law is observed for the latter two RSA classes, while an exponential for the former RSA class.

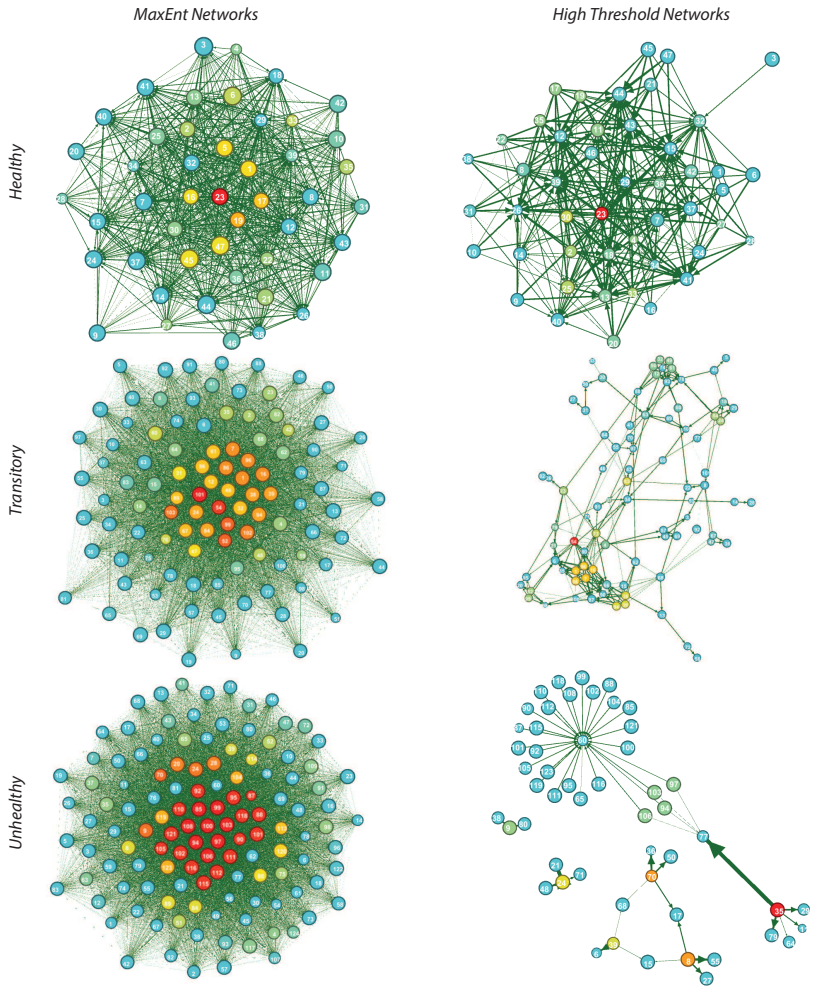

Figure B.3: **Inferred maximum entropy and high-threshold networks.** Maximum entropy microbial networks and high threshold networks are plotted as a function of the microbiome state. Network structure is lost for the transitory and unhealthy microbiome. The color of each node is proportional to the sum of total outgoing TEs of the node (OTE) (the higher OTE, the warmer the color).

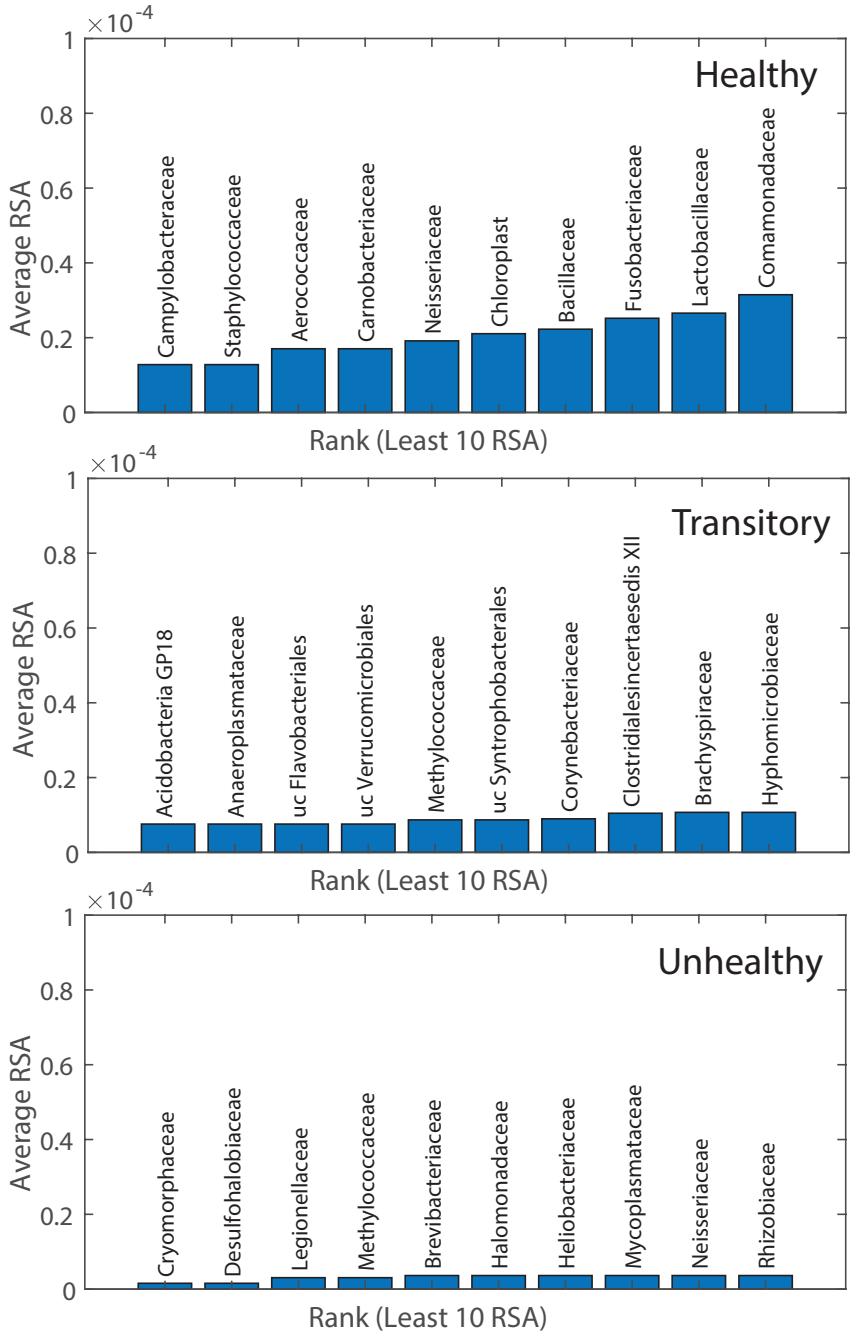

Figure B.4: **Top ten RSA species for each microbiome group.** RSA is reported for the 10 highest RSA species of the healthy, transitory and unhealthy microbiome group. For the unhealthy and healthy group, the top 10 highest RSA species are the most beneficial and detrimental species.

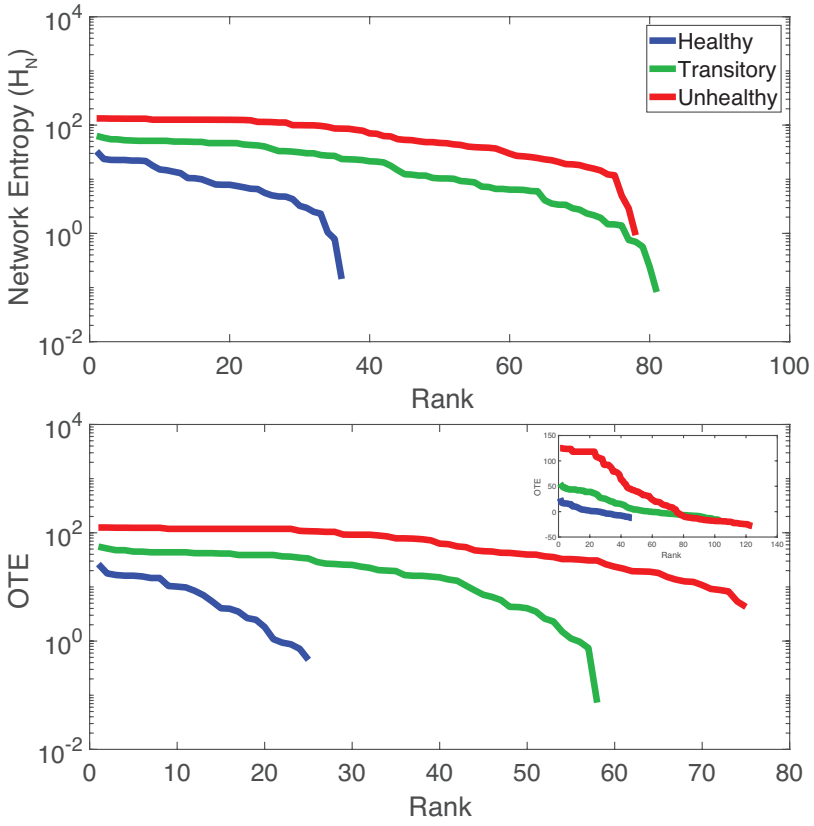

Figure B.5: **Rank-entropy patterns.** The rank of total network entropy and Outgoing Transfer Entropy is plotted in semi-log plots. Many more values of OTE and network entropy are observed for the unhealthy and transitory group.

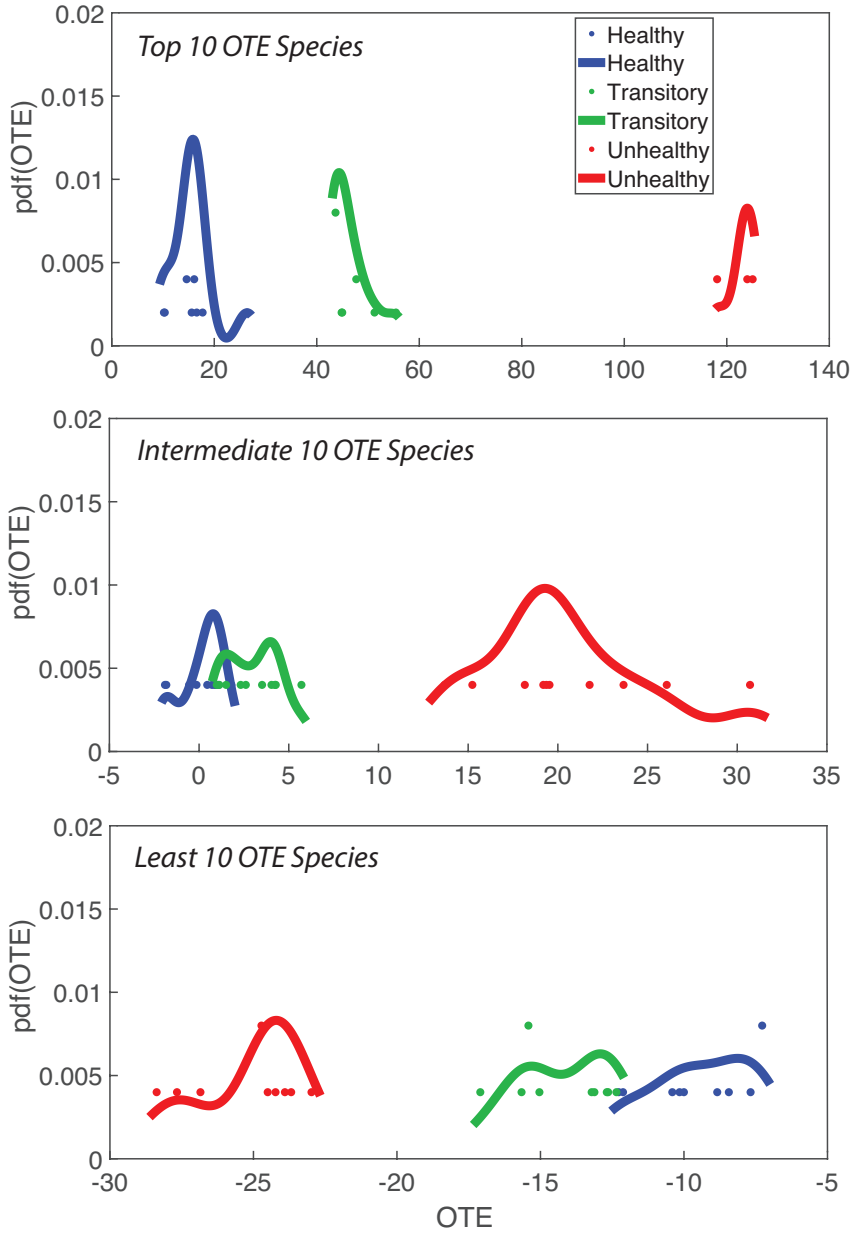

Figure B.6: **Probability distribution function of Outgoing Transfer Entropy.** The top, intermediate and least 10 OTE are plotted considering their probability distribution functions for the healthy, transitory and unhealthy groups. Spline functions fitting the pdfs are shown.

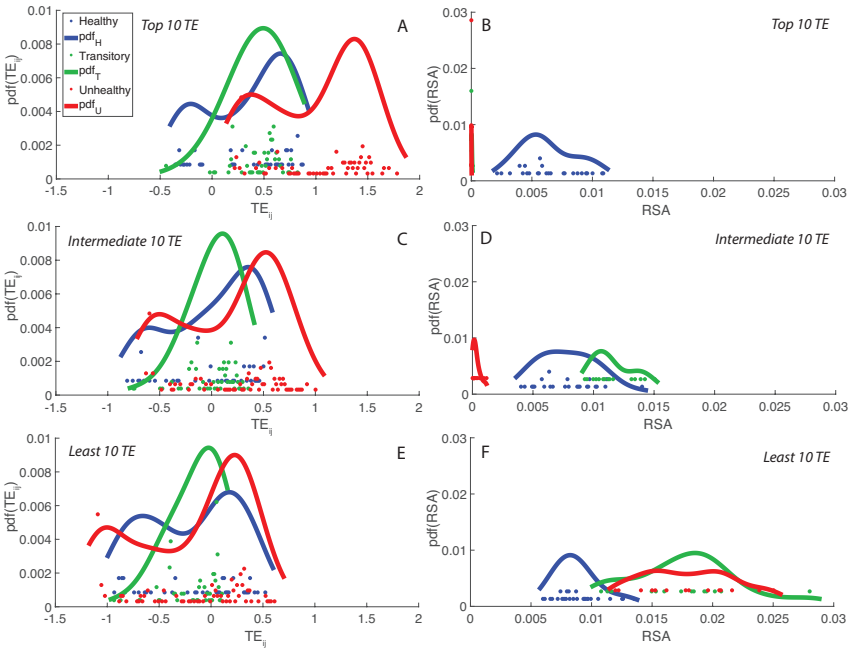

Figure B.7: **Probability distribution function of pairwise Transfer Entropy and RSA.** Pdf for top, intermediate and least 10 pairwise TE and RSA classes are reported as a function of the microbiome group.Spline function fitting of the pdf is shown.

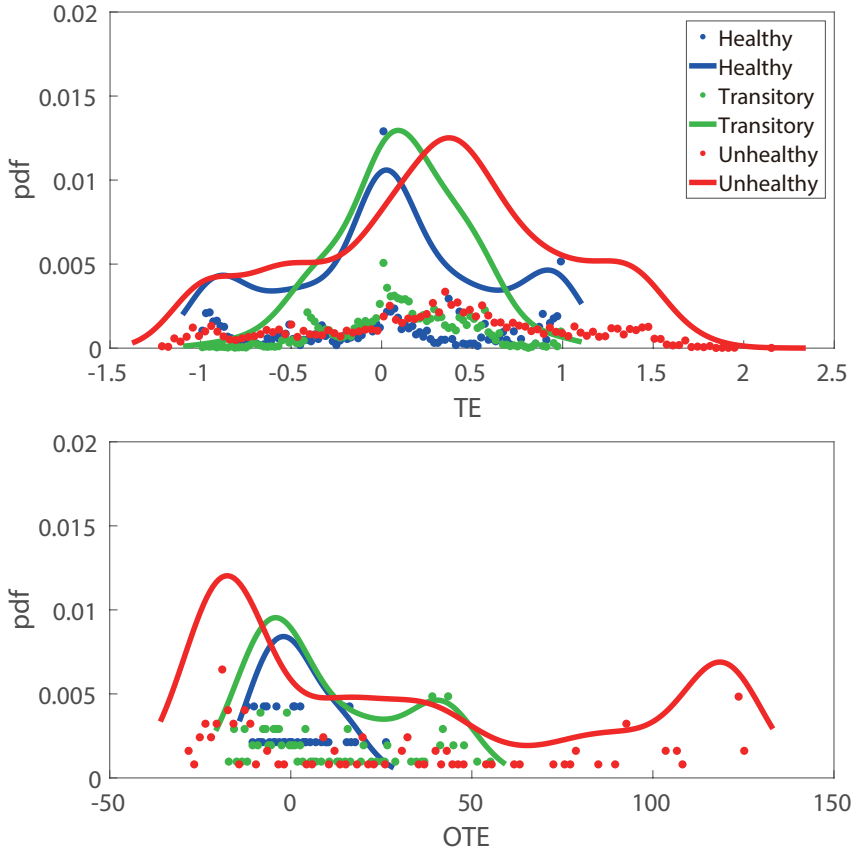

Figure B.8: **Probability distribution function of TE and OTE.** The pdf of TE and OTE (top and bottom plot) are for all individuals in the healthy, transitory and unhealthy groups.

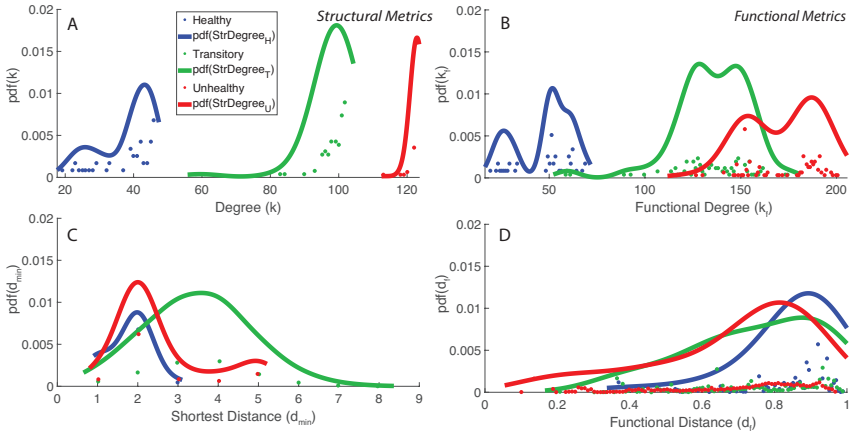

Figure B.9: **Probability distribution of structural and functional microbiome networks.** Pdf of structural and functional network degree and distance are shown on the left and right dependent on the microbiome group. Spline function fitting of pdf is shown.

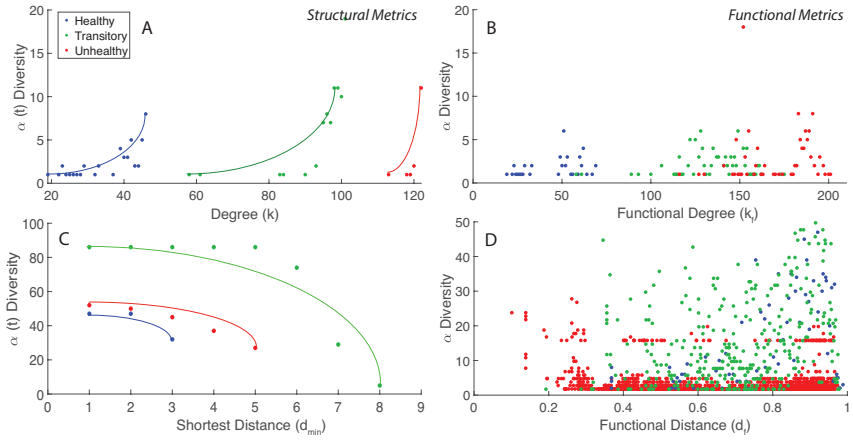

Figure B.10: **Local species diversity as a function of microbiome network features.** Polynomial functions are used to fit the relationship between macroecological indicators and structural network features. Only data are shown for these relationships considering functional network features since no clear fitting function is detected.

Appendix C

Supplement for Chapter 4

Species NO.	**Species Name**	**Fish Stock**	**Native/Invasive**
1	Aurelia.sp	Yes	Native
2	Engraulis.japonicus	Yes	Native
3	Plotosus.lineatus	No	Invasive
4	Sebastes.inermis	Yes	Native
5	Trachurus.japonicus	Yes	Native
6	Girella.punctata	Yes	Native
7	Pseudolabrus.sieboldi	Yes	Native
8	Halichoeres.poecilopterus	Yes	Native
9	Halichoeres.tenuispinnis	No	Native
10	Chaenogobius.gulosus	No	Native
11	Pterogobius.zonoleucus	No	Native
12	Tridentiger.trigonocephalus	No	Native
13	Siganus.fuscescens	Yes	Invasive
14	Sphyraena.pinguis	Yes	Native
15	Rudarius.ercodes	No	Invasive

Table C.1: Species ID considering the Maizuru dataset, scientific and common name, categorization in terms of fish stock, location endemicity (native/invasive).

Species	OTE	ITE	Mean	Std
Aurelia.sp	6.4767	9.612	23.826	126.05
Engraulis.japonicus	8.1583	3.4124	68.056	379.3
Plotosus.lineatus	7.0584	3.1744	28.253	133.03
Sebastes.inermis	6.16	10.524	31.874	50.512
Trachurus.japonicus	6.3931	8.029	174.94	258.7
Girella.punctata	6.0511	7.7665	14.863	29.911
Pseudolabrus.sieboldi	6.2727	10.308	7.3333	6.9701
Halichoeres.poecilopterus	6.1276	6.978	7.5649	11.755
Halichoeres.tenuispinnis	6.6597	5.7889	17.575	33.523
Chaenogobius.gulosus	8.0477	2.5391	8.9386	51.713
Pterogobius.zonoleucus	6.2296	6.8386	18.542	77.554
Tridentiger.trigonocephalus	6.9856	11.227	31.458	43.289
Siganus.fuscescens	7.0709	3.7683	4.6456	25.866
Sphyraena.pinguis	8.1649	2.3686	8.7614	50.779
Rudarius.ercodes	6.0053	9.527	12.142	33.463

Table C.2: Outgoing transfer entropy (OTE), incoming transfer entropy (ITE), mean relative abundance (Mean) and standard deviation (Std).

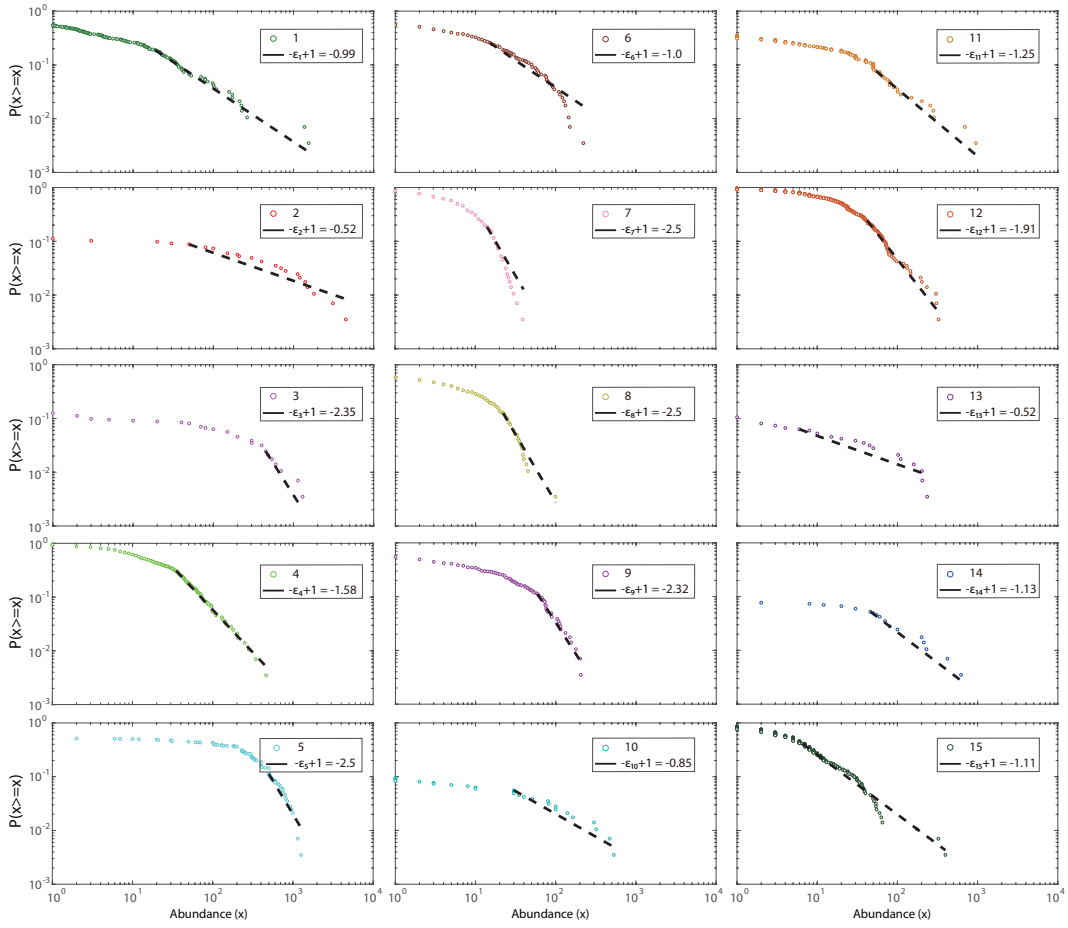

Figure C.1: **EPDF of species abundance.** EPDF of species abundance and power-law fitting.

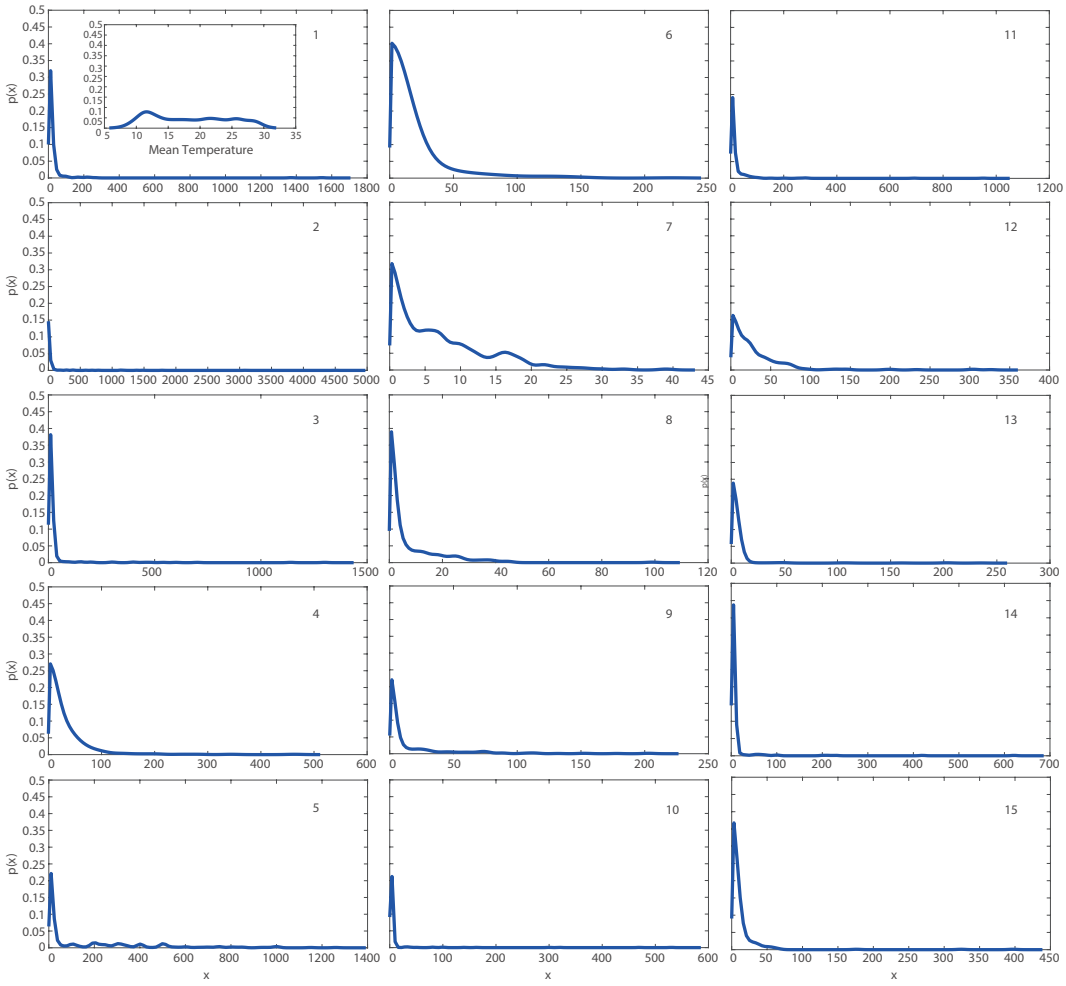

Figure C.2: **Continuous probability distribution function (pdf) of species abundance and mean temperature.**

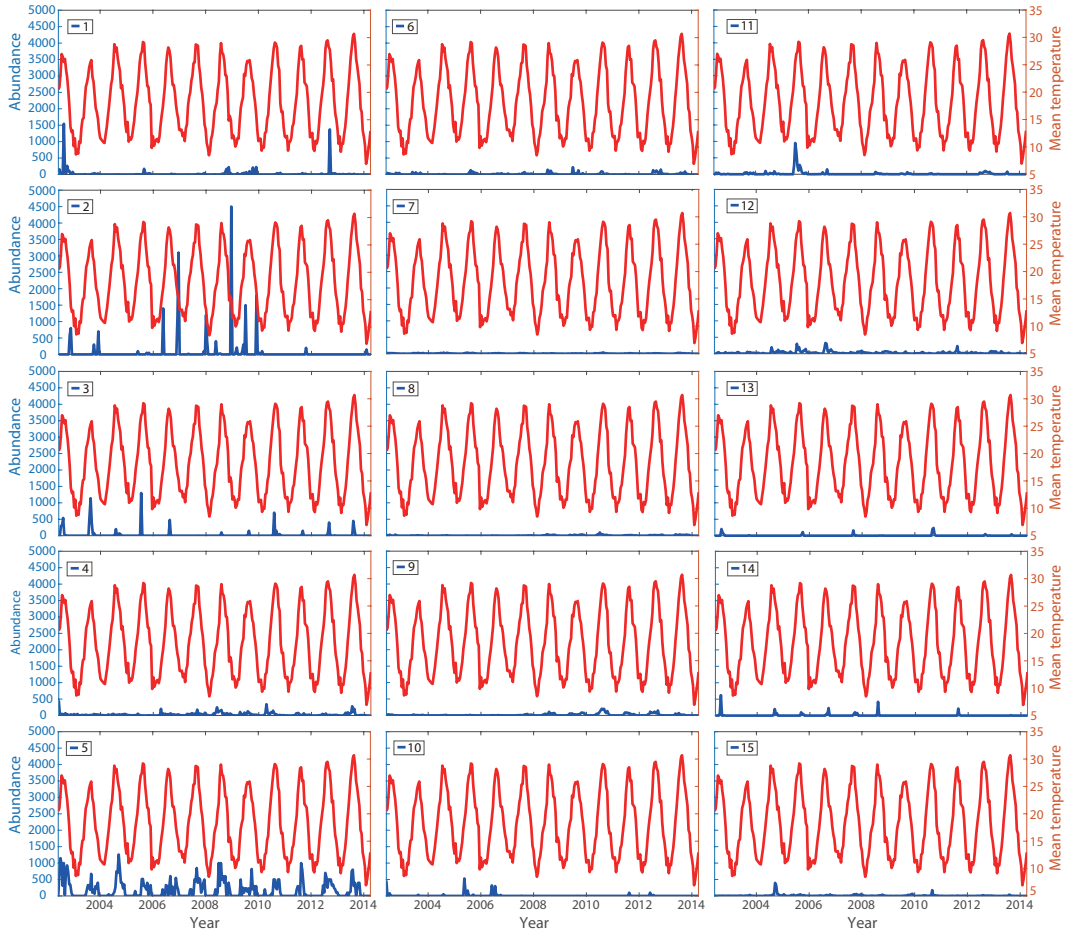

Figure C.3: **Species abundance and mean temperature over time.**

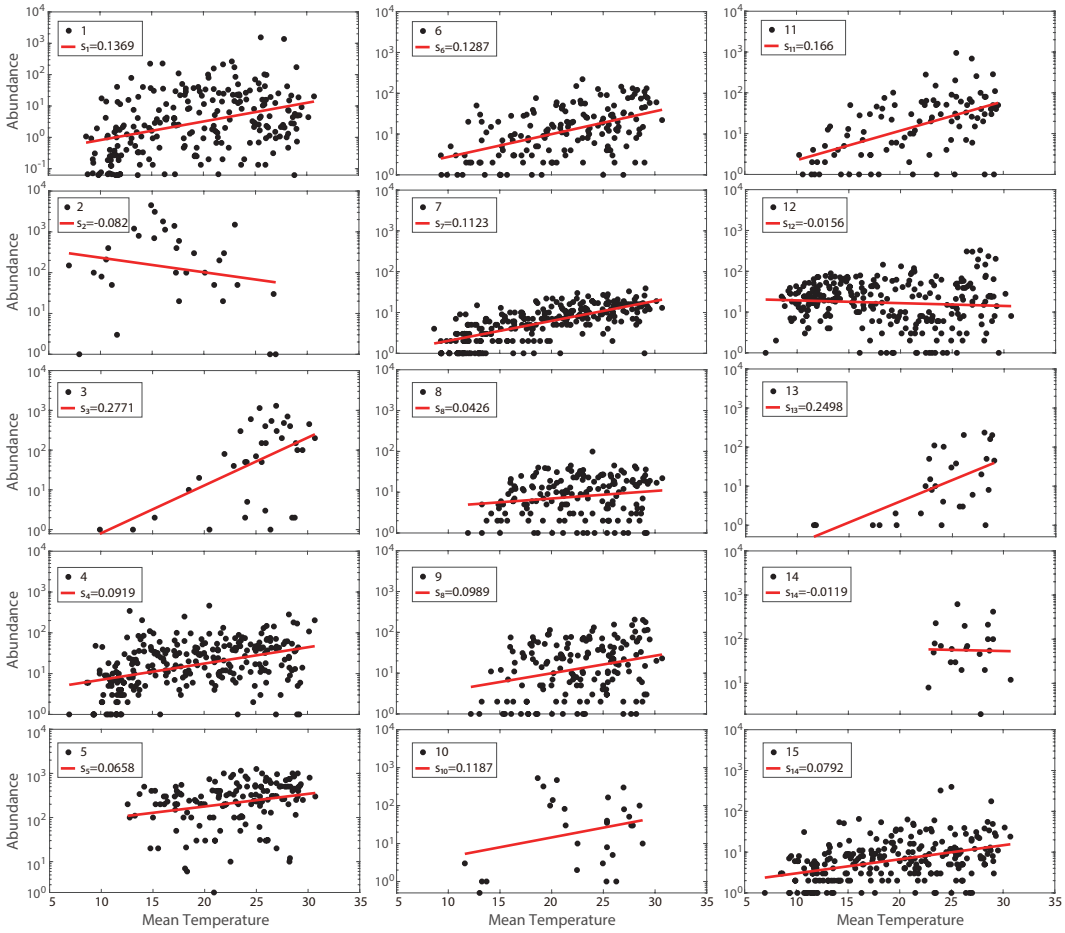

Figure C.4: **The relationship between species abundance and mean temperature.** Species abundance on log scale vs. mean temperature is linearly fitted by the first degree polynomial model (red line).

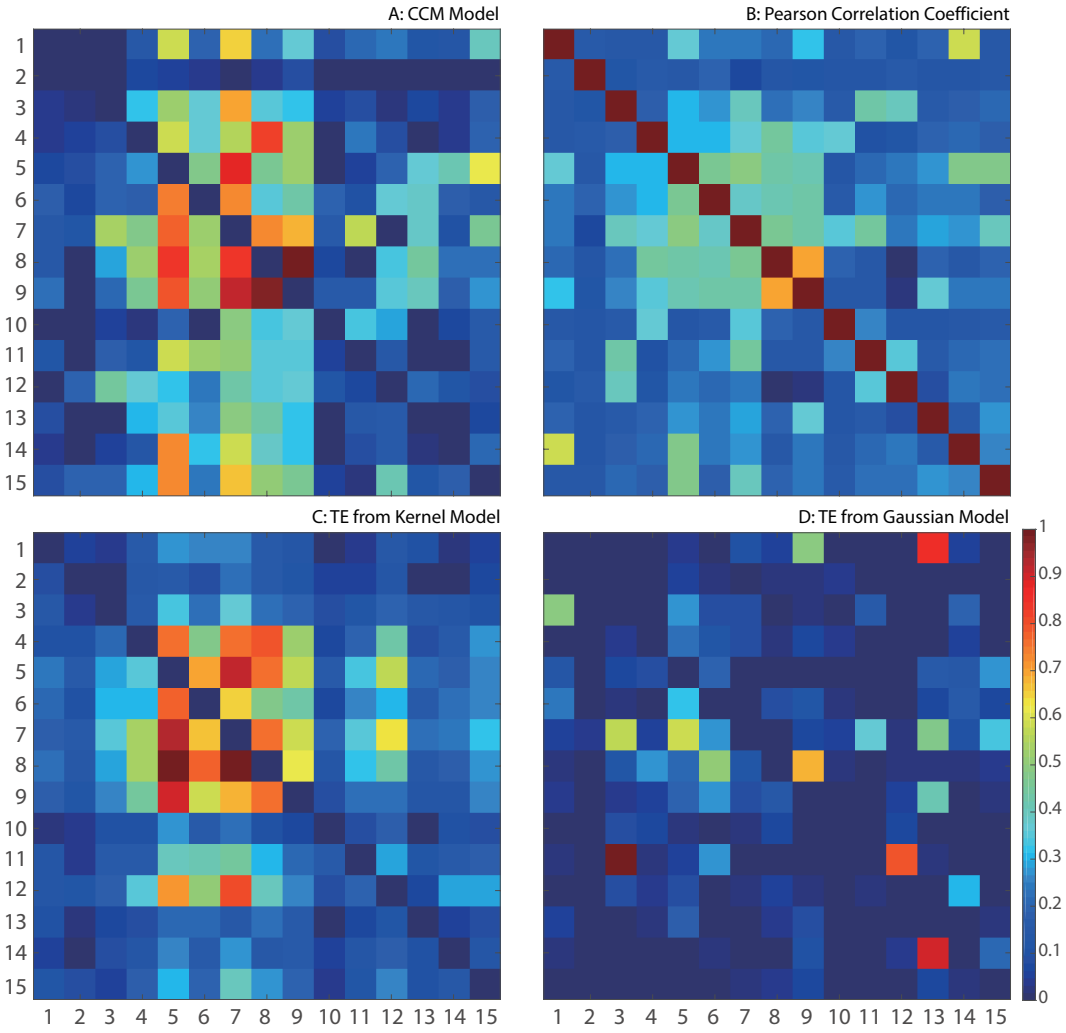

Figure C.5: **Model comparison.** **A**: causal interaction inference from CCM model developed by Sugihara et al. **B**: linear relationship between species computed as Pearson correlation coefficient. **C**: TE-based causal interaction inference using Kernel model, and **D**: Gaussian model.

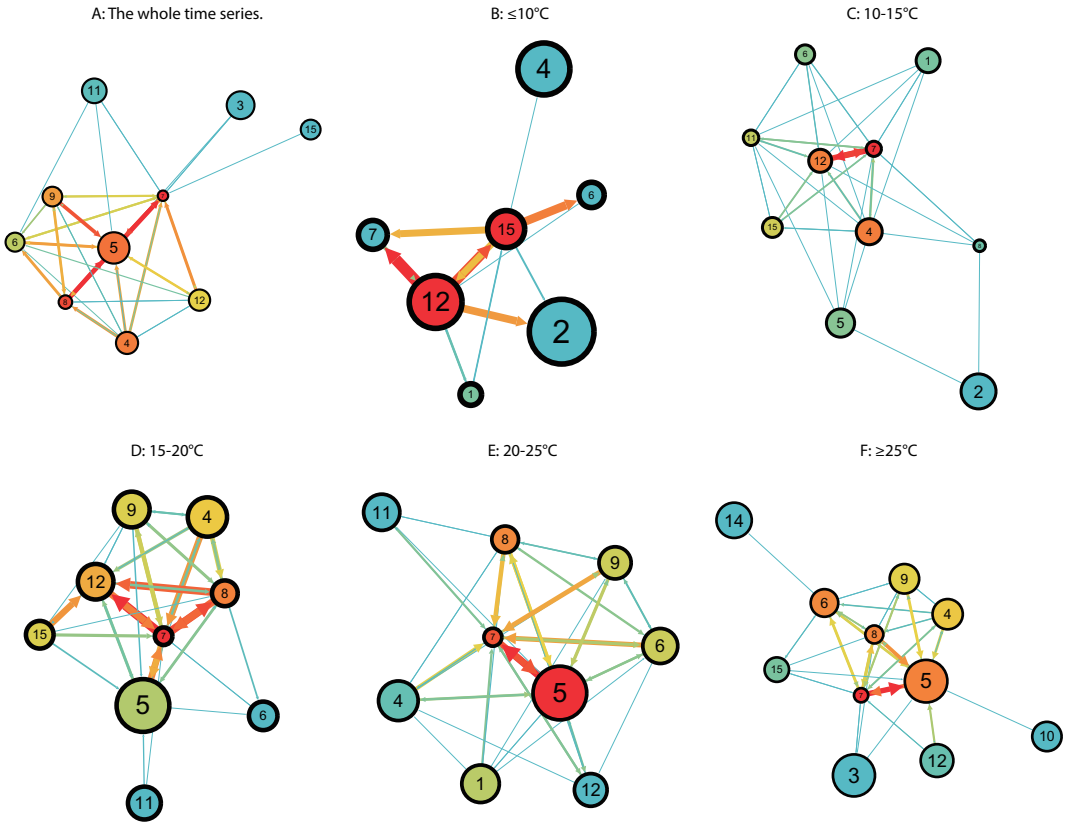

Figure C.6: **OIF-inferred interaction networks of top 20% greatest TEs.** The size of node is proportional to the Shannon Entropy of the species; the color of node is proportional to the total outgoing transfer entropies (OTE) of node (the higher the OTE is, the warmer the node's color is.); the width and color of the link between species are proportional to the TE between the pair of species (The higher the TE is, the warmer (wider) the link's color (width) is.).

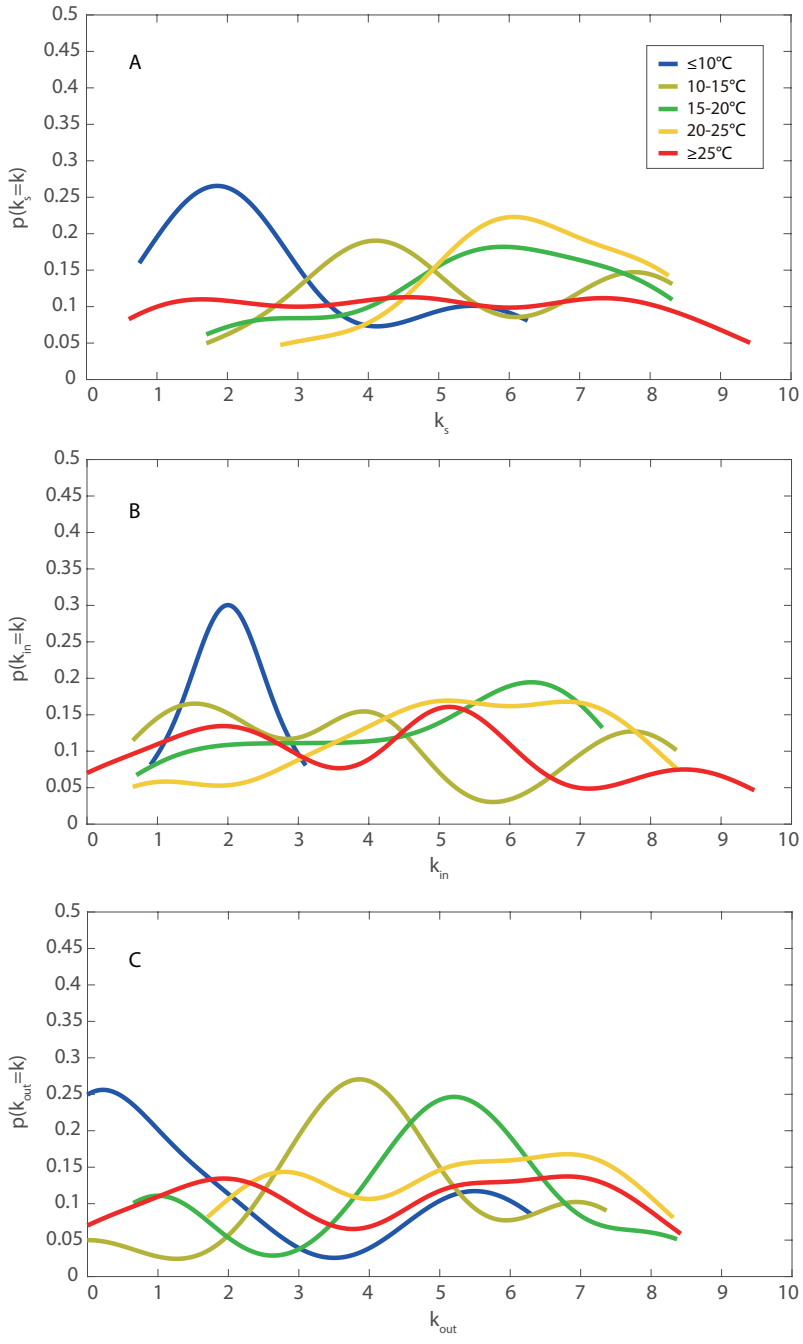

Figure C.7: **Pdfs of nodal degree in OIF networks of top 20% greatest TEs. A**: Pdf of the structural degree, **B**: pdf of the in-degree, **C**: pdf of the out-degree, of nodes in OIF networks corresponding to five MTR groups.

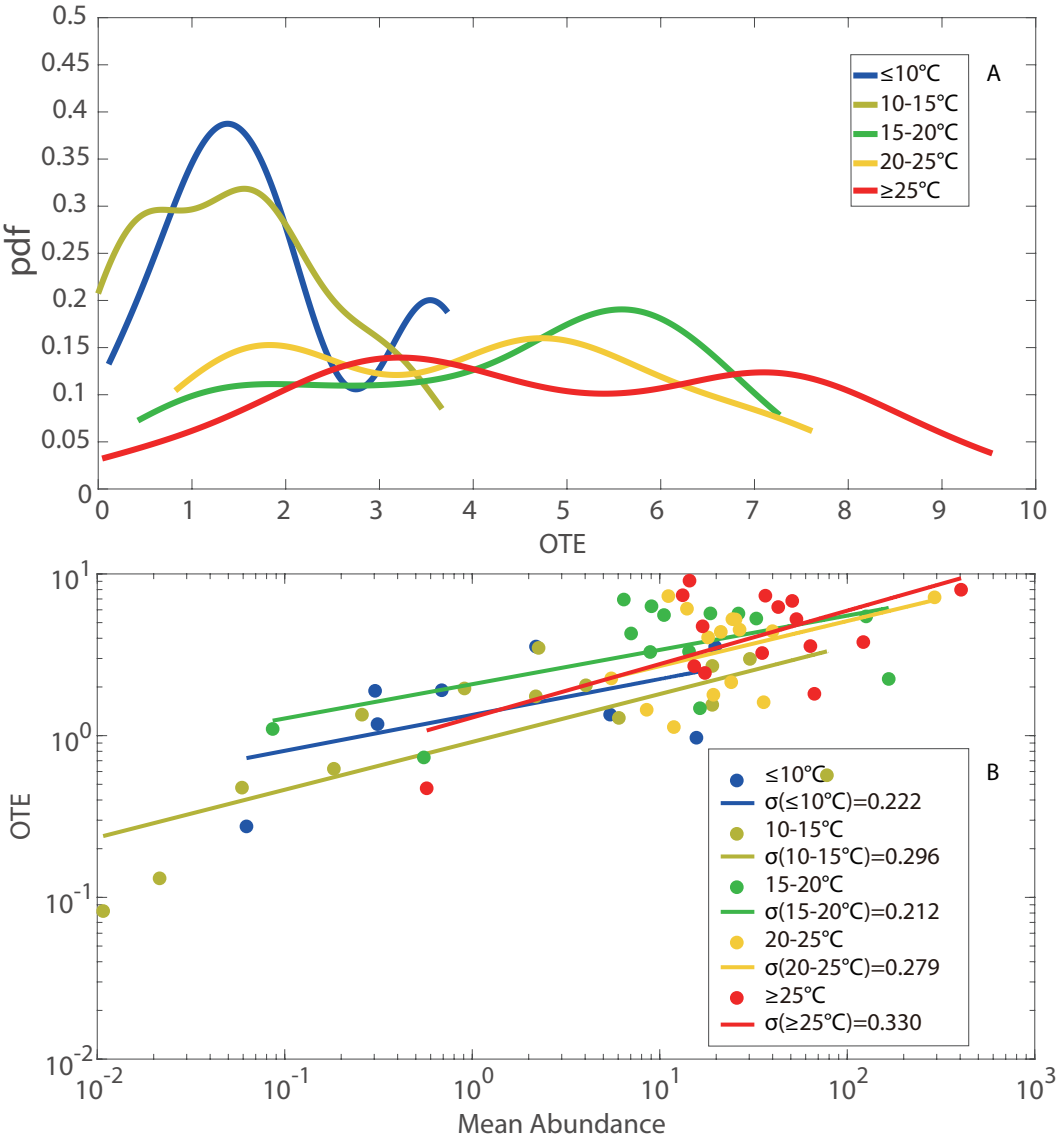

Figure C.8: **Pdf of OTE and OTE vs. species abundance.** **A**: Pdf of OTE of all species for five MTR groups. **B**: OTE vs. mean species abundance on log-log scale illustrates the mutual alteration pattern between the information exchange and species abundance and is fitted by power-law function (straight lines in plot **B**).

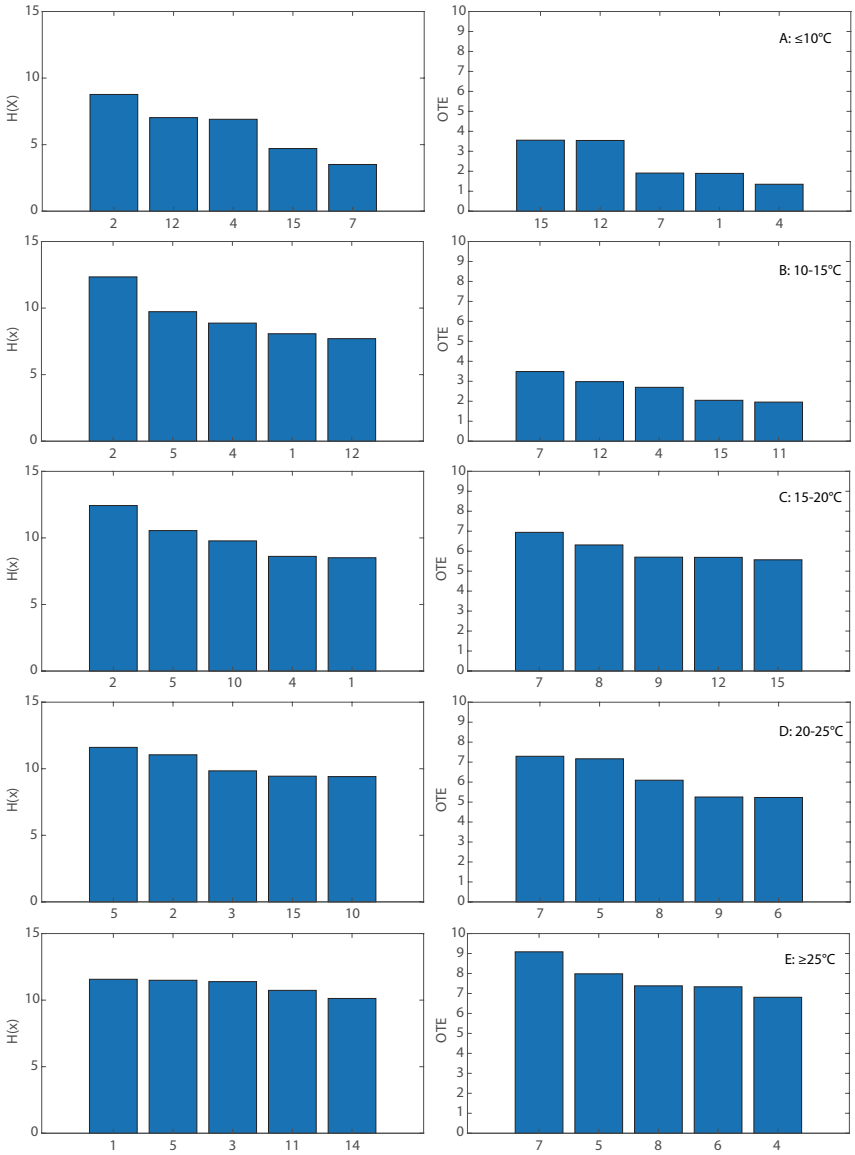

Figure C.9: **Species with top 5 greatest Shannon entropy and influences on other species.** On the **left** plots, species with top 5 greatest information content quantified by Shannon entropy are ranked for five temperature ranges. On the **right** plots, the top 5 most active species in terms of OTE are ranked for five temperature ranges.

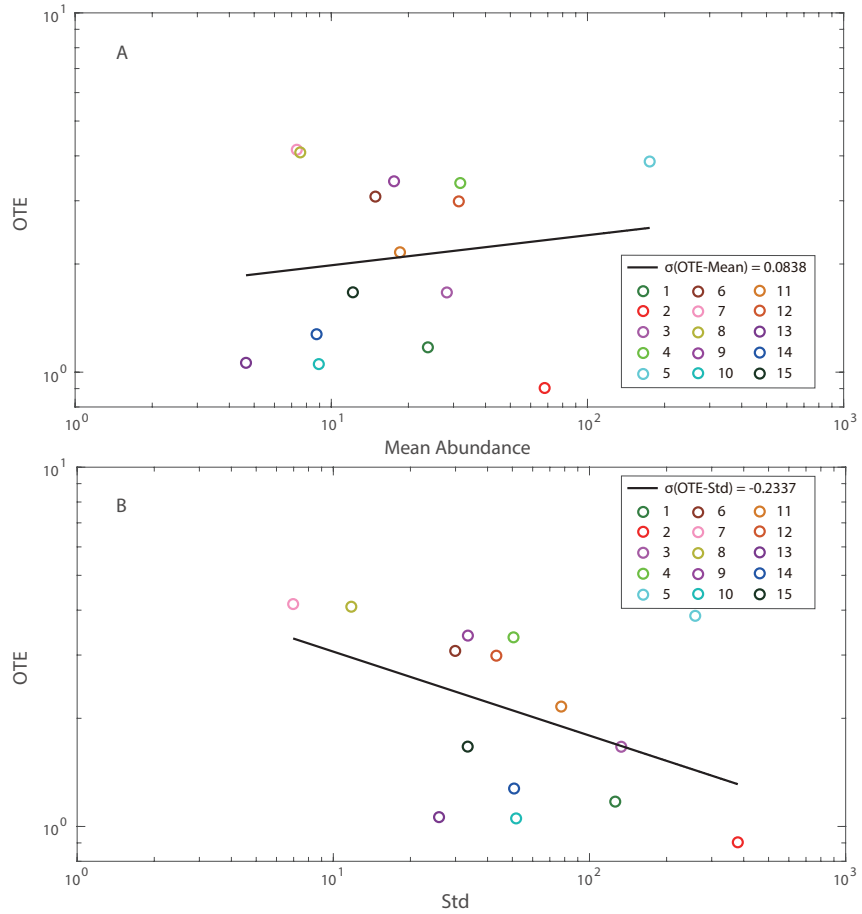

Figure C.10: **OTE against mean and standard deviation of species abundance.**

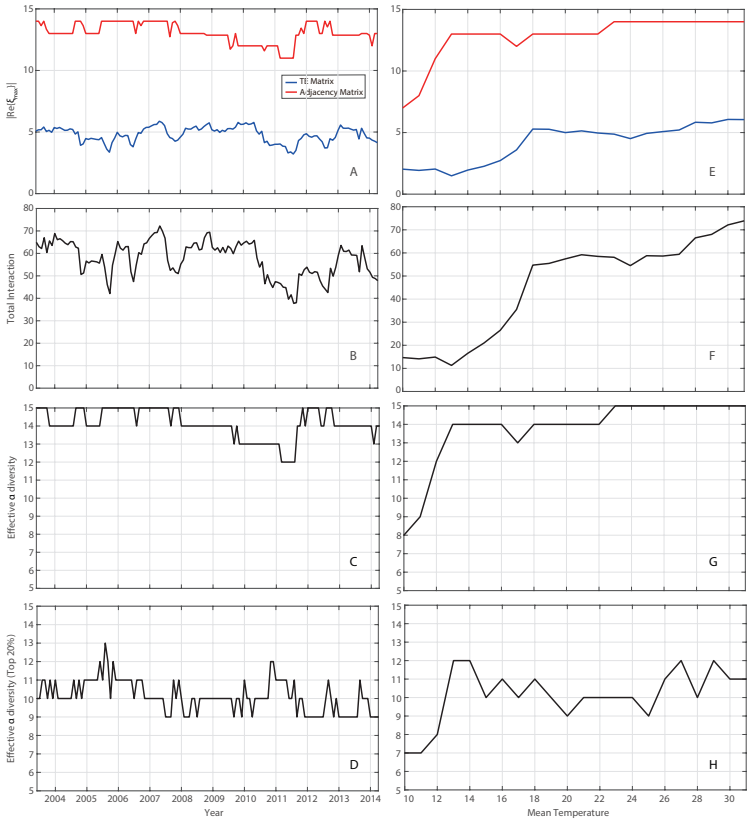

Figure C.11: **Temporally and temperature-dependently dynamical networks.** The stability of the fish ecosystem is indicated as eigenvalues of the TE matrix. Total interactions are calculated as the sum of all TE values in the TE matrix. Effective α diversity is the number of all connected nodes (species) in dynamical networks elaborated from OIF-inferred TE matrix. **A**: the real part of the dominant eigenvalue of temporally dynamical TE matrices (blue line) and corresponding adjacency matrices (red line) over time. **B**: Total interactions of temporally dynamical TE interaction matrices over time. **C**: effective α diversity of temporally dynamical TE interaction matrices without threshold over time. **D**: effective α diversity of temporally dynamical TE interaction matrices with 20% TE threshold over time. **E**: the real part of the dominant eigenvalue of temperature-dependently dynamical TE matrices (blue line) and corresponding adjacency matrices (red line) over mean temperature. **F**: Total interactions of temperature-dependently dynamical TE interaction matrices over mean temperature. **G**: effective α diversity of temperature-dependently dynamical TE interaction matrices without threshold over mean temperature. **H**: effective α diversity of temperature-dependently dynamical TE interaction matrices with 20% TE threshold over mean temperature.

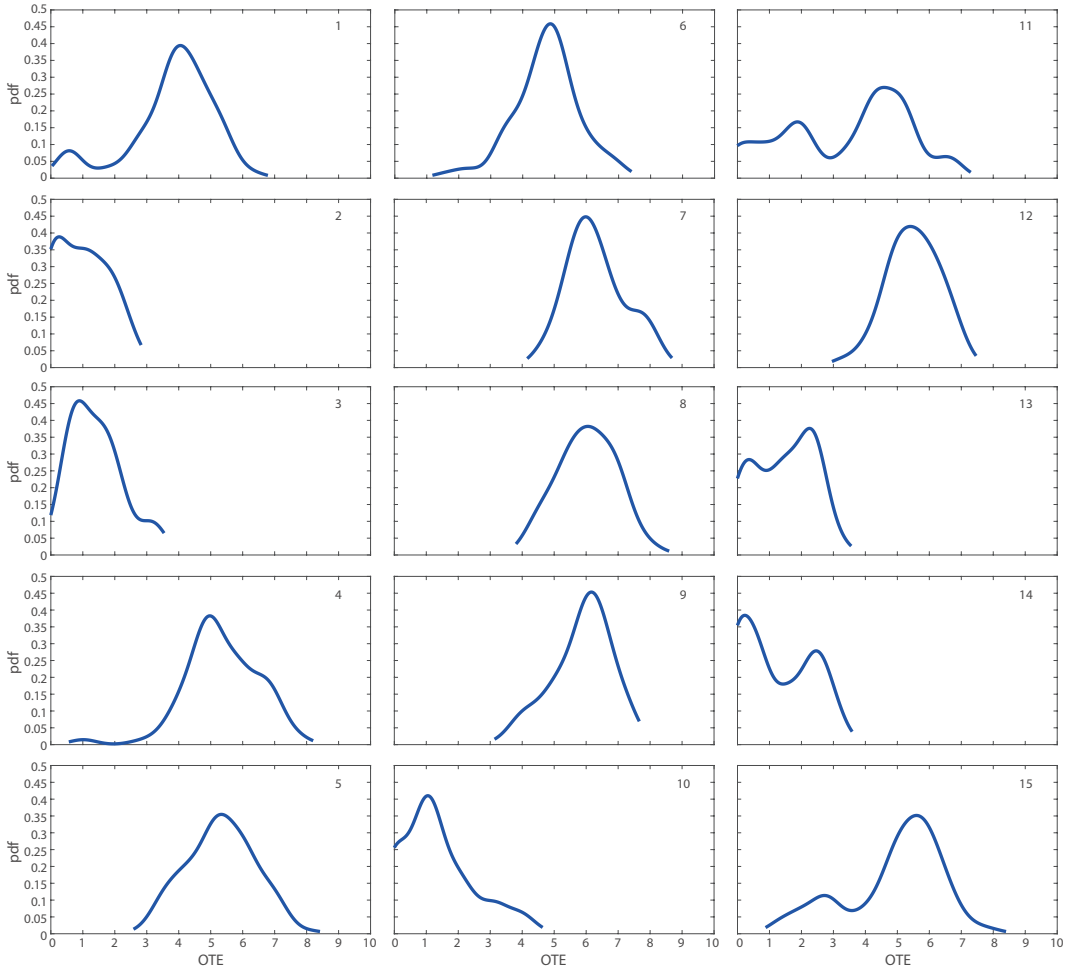

Figure C.12: **Continuous probability distribution function (pdf) of OTE.** Considering the whole time series, TE-based interaction matrix is inferred by OIF model. Pdf of OTE is estimated for all species.