Copulas reveal complex and informative dependencies propagating throughout ecology

Shyamolina Ghosh, Lawrence W. Sheppard, Mark T. Holder, Terrance E. Loecke, Philip C. Reid, James D. Bever, Daniel C. Reuman

Front matter

Affiliations:

Ghosh: Department of Ecology and Evolutionary Biology and Kansas Biological Survey, University of Kansas, Lawrence, KS, 66045, USA

Sheppard: Department of Ecology and Evolutionary Biology and Kansas Biological Survey, University of Kansas, Lawrence, KS, 66045, USA

Holder: Department of Ecology and Evolutionary Biology and Biodiversity Institute, University of Kansas, Lawrence, KS, 66045, USA

Loecke: Environmental Studies Program and Kansas Biological Survey, University of Kansas, Lawrence, KS, 66047, USA

Reid: School of Biological & Marine Sciences, University of Plymouth, Drake Circus, Plymouth PL4 8AA, UK; Continuous Plankton Recorder Survey, The Marine Biological Association, The Laboratory, Citadel Hill, Plymouth PL1 2PB, UK

Bever: Department of Ecology and Evolutionary Biology and Kansas Biological Survey, University of Kansas, Lawrence, KS, 66045, USA

Reuman: Department of Ecology and Evolutionary Biology and Kansas Biological Survey, University of Kansas, Lawrence, KS, 66047, USA; Laboratory of Populations, Rockefeller University, 1230 York Ave, New York, NY, 10065, USA

Keywords: dependence, regression, correlation, copula, population, community

Type of Article: Ideas and Perspectives

The number of words in the abstract: 250

The number of words in the main text (excluding abstract, acknowledgements, references, table and figure legends): 7497

The number of words in each text box: No text boxes

The number of references: 34

The number of figures, tables, and text boxes: 7 figures, 3 tables, 0 text boxes

The name and complete mailing address (including telephone and fax numbers and e-mail address) of the person to whom correspondence should be sent: Daniel Reuman, 2101 Constant Ave, Lawrence, KS, 66047, reuman@ku.edu, 626 560 7084.

Abstract

All branches of ecology study relationships among environmental and biological variables. However, ubiquitously used approaches to studying such relationships, based on correlation and regression, provide only a small slice of the complex information contained in the relationships. Other statistical approaches exist that provide a complete description of relationships between variables, based on the concept of the *copula*; they are applied in finance, neuroscience and other fields, but not in ecology. We here explore the concepts that underpin copulas and examine the potential for those concepts to improve our understanding of ecology. We find that informative copula structure in dependencies between variables is exceedingly common across all the environmental, species-trait, phenological, population, community, and ecosystem functioning datasets we considered. Many datasets exhibited asymmetric tail dependence, whereby two variables are more strongly related in their left compared to right tails, or vice versa. We describe mechanisms by which observed copula structure and tail dependence can arise in ecological data, including a Moran-like effect whereby dependence structures between environmental variables are inherited by ecological variables; and asymmetric or nonlinear influences of environments on ecological variables, such as under Liebig's law of the minimum. Copula structure and tail dependence can also reveal causal relationships between variables. We describe consequences of copula structure, including impacts on extinction risk and the stability through time of ecosystem services. By documenting the importance of a complete description of dependence, advancing conceptual frameworks, and demonstrating a powerful approach, we aim to elevate copula reasoning to widespread use in ecology.

Introduction

All branches of ecology study relationships among environmental and biological variables. However, commonly used correlation and regression approaches to studying such relationships are limited, and provide only a partial description of a relationship between variables. For instance, datasets showing markedly different relationships may have the same correlation coefficient (Fig. 1A-B). The variables of Fig. 1A (respectively, Fig. 1B) are principally related in the left (respectively, right) portions of their distributions, an asymmetric pattern of dependence that can have profound significance, as discussed below, but is not captured by correlation. Monotonic transformations can alter the Pearson, but not the Spearman correlation (Appendix S1, Fig. S1). Although correlation is only one way to study relationships between variables, other common metrics also provide partial information.

Well-developed approaches do exist, however, and are applied widely in other fields, that provide a description of the dependence between two or more variables that can be mathematically proven to be complete (Nelsen 2006; Joe 2014). The approaches are based on the concept of the *copula*, which we here introduce. Copula approaches separate the information of a bivariate random variable into two distinct parts: the information in the marginal distributions (which says nothing about the dependence between the variables), and the remaining information (which is solely about dependence). From a data perspective, given a sample (x_i, y_i), $i = 1, 2, \ldots, n$ (e.g., Fig. 1A-B), the dependence structure is revealed by the bivariate plot of u_i versus v_i, where u_i is the rank of x_i in the set x_j $(j = 1, \ldots, n)$, divided by $n + 1$; and v_i is defined in the same way but using the y_i. The rank of the smallest element in a set is 1. The u_i and v_i are called *normalized ranks* of the x_i and y_i. See, for instance, Fig. 1C-D (the points), which show the copula structures of A-B, respectively. Ranking makes the marginal distributions of the underlying random variables uniform, isolating the dependence structure. See Nelsen (2006) and Joe (2014) for general background. A theorem of Sklar (1959) states that copulas give a unique and complete representation of the dependence between variables.

A main benefit of a copula approach is that it can sensitively detect tail dependence and asymmetries of tail dependence. Tail dependence is association between extreme values of variables. If smaller values of two positively associated variables are more strongly associated than are larger values, the variables are said to have stronger lower- or left-tail dependence than right- or upper-tail dependence (Fig. 1A, C); and *vice versa* if larger values are more strongly associated than smaller ones (Fig. 1B, D). Datasets of the same correlation can have a range of tail dependence (Fig. S2), which can be quantified (see section on methods for the observational hypothesis).

The goal of this paper is to illustrate the potential for applications of copulas in ecology. Our results provide initial support for the following overall hypothesis: environmental, ecological and evolutionary processes generate and transmit copula structure and tail dependence, with substantial consequences for theoretical and conceptual understanding, for inference, and for applications across ecology. This is a broad hypothesis, and even the substantial empirical and theoretical evidence we present here has only begun to test it. So we hope our work stimulates further evaluation of the hypothesis by

the research community. We examine our overall hypothesis by assessing several more specific hypotheses, described below and outlined in Fig. 2, all of which bear upon the overall hypothesis.

One way we may expect a copula approach to better reflect ecological relationships relates to Liebig's law of the minimum. If, for instance, the growth of a plant depends on soil nitrogen, N, and other factors, a plot of plant growth rates versus N may look more like Fig. 1A or C than like Fig. 1B or D. Nitrogen will control plant growth, producing a clear relationship, only when it is limiting, in the left portions of the distributions. The portion of the distributions that dominate the association is visible in the copula approach, but not via standard techniques. We therefore frame "causal hypothesis" A (labeled in Fig. 2): copulas of dependencies between environmental and ecological variables will sensitively reveal nonlinear, asymmetric and threshold effects for population, community, and ecosystem functioning variables; copulas between environmental and ecological variables will commonly show tail dependence; such tail dependence will lead to tail dependence across space in ecological variables.

A second reason we may expect a copula approach to benefit ecological research is that prior work has demonstrated important copula stucture in the spatial dependence of environmental variables (Serinaldi 2008; Li *et al.* 2013; Goswami *et al.* 2018; She & Xia 2018). One may expect copula structure in environmental variables to take the form of strong tail dependence if intense meteorological events are also widespread in their effects, as seems frequently to be the case. Spatial dependence in an environmental variable tends to beget spatial dependence in fluctuations of populations or other ecosystem variables (this is called *spatial synchrony*) which are influenced by the environmental variable. This is called the *Moran effect*. If copula structure is transmitted, possibly in modified form, from environmental to ecological variables, then we may expect important copula structure to be a common feature of the spatial synchrony of population, community, biogeochemical and other ecological variables. Synchrony is a phenomenon of great interest in ecology (Liebhold *et al.* 2004; Sheppard *et al.* 2016; Walter *et al.* 2017). We therefore frame causal hypothesis B (again labeled in Fig. 2): Moran effects transmit copula structure in spatial dependence from environmental to ecological variables, and hence copula structure and tail dependence will be common features of spatial synchrony in ecological variables.

A third reason we may expect a copula approach to benefit ecological research is that such an approach may help illuminate the causal mechanisms between variables. For instance, if two species Sp1 and Sp2 are strong competitors, abundances of the two species across quadrats should be negatively related, as in Fig. 1E-F. If Sp1 is plotted on the horizontal axis and Sp2 on the vertical, then Fig. 1E would suggest Sp1 is the dominant competitor: when Sp1 is abundant, Sp2 is necessarily rare because it is suppressed; whereas when Sp1 is rare, Sp2 may be abundant, or may also be relatively rare due to limiting factors other than Sp1. Alternatively, Fig. 1F suggests that Sp2 is dominant. The copula approach again reveals which portions of the distributions are principally responsible for the association between variables, and this information can help support or refute alternative causal hypotheses. We frame causal hypothesis C (Fig. 2): copulas of abundances through time or space of two competing or otherwise interacting species will reflect any asymmetries in the

interaction. We will later frame causal hypothesis D, which concerns character evolution.

Another reason we may expect copula structure to be important in ecology has to do with spatially aggregated or averaged quantities. Many ecological variables of applied importance depend on the spatial average or sum of local quantities. For instance, regional methane and CO_2 fluxes are the sum of local fluxes; and the total economic value of a fishery is the sum of local catches. We therefore frame "consequential hypothesis" X (labeled on Fig. 2): details, revealed by copulas, of the dependence between local quantities can influence the spatial mean or sum, and this may influence higher organizational levels in ecology and human concerns. Illustrating the idea here, prior work demonstrated that the overall reliability of wind-generated electricity depends sensitively on details of the dependence between wind speeds at multiple generator sites (Li *et al.* 2013). Spatially aggregated ecological variables may show a similar dependence, with potentially substantial consequences. For instance, if populations of a pest species in different locations are all positively associated and also strongly related to each other in their right tails, then local outbreaks will tend to occur together, creating regional epidemics. On the other hand left-tail dependence in a pest species, even if overall correlation were unchanged, would have more benign effects. We will later frame consequential hypotheses Y, about extinction risk, and Z, about Taylor's law.

Our causal and consequential hypotheses imply, or rely on, an "observational hypothesis" (O in Fig. 2), which we here frame. Datasets in ecology will commonly have copula structure distinct from that of a Gaussian/normal distribution (here called non-normal copula structure); they will commonly show tail dependence distinct from that of a normal distribution; in particular they will show asymmetric tail dependence.

Because the observational hypothesis bears upon all the causal and consequential hypotheses, we test it first, in sections 2-4 of the paper. We analyze several datasets, including environmental, species-trait, phenological, population, community, and ecosystem functioning data, selected to span multiple sub-fields and organizational levels of ecology. Sections 5 and 6 test the causal hypotheses, and sections 7 and 8 test the consequential hypotheses. Section 9 is the Discussion. Copula approaches have been used to great effect in neuroscience (Onken *et al.* 2009), bioinformatics (Kim *et al.* 2008), medical research (Emura & Chen 2016), direct study of environmental variables (Serinaldi 2008; Li *et al.* 2013; Goswami *et al.* 2018; She & Xia 2018), and finance (Li 2000). To our knowledge, these concepts have never been used in ecology. We argue that benefits to ecology of using a complete copula description of dependence will be substantial.

Data

Datasets included environmental, species-trait, phenological, population, community, and ecosystem functioning data (Table 1). Copula structure describing the relationship between atmospheric weather variables such as rainfall or wind speed as measured in multiple locations has been examined previously in the meteorological literature (e.g., Serinaldi 2008; Li *et al.* 2013). We therefore examined environmental variables from the soil instead, using the Rapid Carbon Assessment

database (RaCA; Wills *et al.* 2014). The database comprises measurements of soil organic carbon and total soil nitrogen (Mg C or N per hectare of soil surface) from 5907 locations across the coterminous United States (Fig. S3 for sampling locations). Species-trait data were average species basal metabolic rate (BMR, KJ per hour) and body mass (grams) for 533 species of birds (McNab 2009) and 638 species of mammals (McNab 2008). These data have been much studied, but to our knowledge the copula structure of the dependence between metabolic rate and body mass has not been examined. Species-trait data such as these reflect the coevolution of the two traits. Phenological data were aphid first-flight dates from 10 locations (Fig. S4) across the United Kingdom (UK) for 20 aphid species (Table S1), for the 35 years 1976 to 2010. These time series were computed from the Rothamsted Insect Survey (RIS) suction-trap dataset (Harrington 2014; Bell *et al.* 2015). The first of our two population-level datasets was also derived from the RIS suction-trap data, and comprised total counts of aphids trapped for the same locations, species, and years. Our second population dataset comprised average annual plankton abundance estimates for 14 locations (Fig. S5) in the North Sea and British seas for 22 taxa (Table S2) for the 56 years 1958 to 2013. The locations are 2° by 2° grid cells. Community-level data, obtained from the Cedar Creek Ecosystem Science Reserve, were plant aboveground biomass (Tilman 2018a) and Shannon's diversity index (computed from plant species percent cover data (Tilman 2018b)) for 168 plots (Tilman 2018c), each 9m by 9m, as sampled in the years 1996-2000 and 2007, each year analyzed separately (Tilman *et al.* 2001, 2006). Finally, ecosystem functioning data were methane (CH_4) fluxes between the soil or water surface and the troposphere, measured at 13 locations at daily to weekly intervals from September 2015 to September 2016 at the Great Miami Wetland mitigation bank in Trotwood, Ohio (Holland *et al.* 1999; Jarecke *et al.* 2016; Smyth *et al.* 2019). Each included location was measured on at least 50 dates. See Appendix S2 for additional data details.

The environmental, species-trait, and community datasets are henceforth called the "bivariate" datasets in the sense that they comprise two quantities measured at different locations or for different species (Table 1). The phenology, population, and ecosystem functioning datasets are "multivariate" in that they comprise, for each taxon, measurements through time at multiple locations; copula structure can be studied for each pair of locations.

Background, concepts and methods for testing the observational hypothesis

Conceptually, a bivariate copula is a random variable on the unit square, $[0,1] \times [0,1]$, with uniform maginals. If (X, Y) is a continuous random variable and the cummulative distribution functions of X and Y are $F_X(x) = P[X \leq x]$ and $F_Y(y) = P[Y \leq y]$, then the dependence structure of (X, Y) is revealed by the random variable $(F_X(X), F_Y(Y))$, which takes values in $[0,1] \times [0,1]$ and has uniform marginals.

We introduce the copula families we used, and measures of tail dependence. The normal copula family (henceforth N) is a one-parameter family, with parameter corresponding to the correlation of the related bivariate normal distribtion, and

controlling the degree of dependence (Fig. 1G for an example). Left- and right-tail dependence are 0 for the N copula, and this serves as a benchmark for tail dependence in other copulas - positive values of our tail dependence measures correspond to dependence stronger than an N copula. The Clayton and survival-Clayton families (Fig. 1C-D) are two more one-parameter families. The Clayton has positive lower-tail dependence (0.87 for the example of Fig. 1C) and 0 upper-tail dependence; the survival-Clayton is a rotation of the Clayton by 180° and has positive upper-tail (0.87 in Fig. 1D) and 0 lower-tail dependence. For our model selection procedure (see below), we used several families of copulas exhibiting various lower- and upper-tail dependence characteristics (Clayton, survival-Clayton, Gumbel, survival-Gumbel, Joe, survival-Joe, Frank, BB1, survival-BB1, BB6, survival-BB6, BB7, survival-BB7, BB8, survival-BB8, henceforth abbreviated C, SC, G, SG, J, SJ, F, BB1, SBB1, BB6, SBB6, BB7, SBB7, BB8, SBB8). Copulas are implemented in the R packages `copula` and `VineCopula`. Visualizations are in Figs S6-S21. If LT and UT denote lower- and upper-tail dependence measures, their difference measures asymmetry of tail dependence. See Nelsen (2006), Joe (2014) for further background.

We tested the observational hypothesis (Fig. 2, O) principally via a model selection procedure in which the copulas listed above were fitted to our ecological datasets and fits were compared via the Akaike and Bayesian Information Criteria (AIC and BIC). For bivariate datasets (x_i, y_i) for $i = 1, ..., n$, model selection involved several steps. First, we produced normalized ranks u_i and v_i as in the Introduction. Second, we tested the independence of the u_i and v_i (using `BiCopIndTest` in the `VineCopula` package). Our model selection algorithms were ineffective if data could not be distinguished from independent data since many of the copula families we considered include the independent copula. If independence could be rejected (0.05 significance level), model selection proceeded. Third, we fit the 16 bivariate copula families listed previously to the normalized ranks via maximum likelihood, using `BiCopEst` in `VineCopula`, then obtaining AIC and BIC values and accompanying model weights $\mathrm{AIC_w}$ and $\mathrm{BIC_w}$ (Burnham & Anderson 2003). `BiCopEst` also provided the lower- and upper-tail dependence of the best-fitting member of each family. Fourth, $\mathrm{AIC_w}$ values were used to get model-averaged lower- and upper-tail dependence values using standard model averaging formulas (Burnham & Anderson 2003); likewise for BIC. Model selection methods give the *relative* support of several models, but do not indicate whether any of the models are an objectively good fit. To test that, and as our final step, we tested the goodness of fit of our AIC-best copula family using a bootstrapping procedure (`BiCopGofTest` of `VineCopula`). The procedure performed one test based on a Cramer-von Mises statistic and another based on a Kolmogorov-Smirnov statistic. To keep computation times reasonable, a run using 100 bootstraps was performed; if the p-value from either test was less than 0.2, tests were re-run with 1000 boostraps.

The first part of the observational hypothesis was tested for a bivariate dataset by comparing the AIC weights of the non-N copula families with that of the N family. The second part was tested by comparing the model-averaged lower- and upper-tail dependence values to the lower- and upper-tail dependence values of an N copula (i.e., 0). The third part was tested by comparing the difference of the model-averaged lower- and upper-tail dependence values to the analogous

difference for an N copula, again 0. For multivariate datasets, we performed the bivariate analysis for all possible pairwise combinations of locations, and compared AIC weights and tail-dependence of minimum-AIC copulas with that of an N copula. We also counted the number of location pairs for which model-averaged lower-tail dependence was greater than model-averaged upper-tail dependence or vice-versa, and reported the percentage of non-independent location pairs showing stronger lower-tail dependence.

In addition to our model selection approach, we also used nonparametric approaches to provide greater confidence in our tests of the observational hypothesis. We used three statistics which quantify the extent to which the normalized ranks u_i and v_i are related in any part of their distributions; we here describe the statistics conceptually, with details in Appendix S3. The statistics are defined with positively associated variables in mind. All our variables were positively associated when they were significantly associated. Given two bounds $0 \leq l_b < u_b \leq 1$, define the lines $u + v = 2l_b$ and $u + v = 2u_b$, which intersect the unit square (Figs 1H, S22). Our statistics quantify the dependence between u_i and v_i in the region bounded by these lines. Using $l_b = 0$ and $u_b \leq 0.5$ gives information about dependence in the left parts of the distributions of u and v, and using $u_b = 1$ and $l_b \geq 0.5$ gives information about dependence in the right parts. The first statistic, $\mathrm{cor}_{l_b,u_b}(u, v)$, is the portion of the Spearman correlation of u_i and v_i that is attributable to the points between the bounds. Larger values indicate stronger positive association. We also defined a statistic P_{l_b,u_b} (Appendix S3), which has a similar interpretation to $\mathrm{cor}_{l_b,u_b}(u, v)$. Our third statistic, $\mathrm{D}^2_{l_b,u_b}(u, v)$, is the average squared distance between points satisfying $u_i + v_i > 2l_b$ and $u_i + v_i < 2u_b$ and the line $v = u$. Unlike cor_{l_b,u_b} and P_{l_b,u_b}, for which large values inidicate strong dependence between the bounds, small values of $\mathrm{D}^2_{l_b,u_b}$ indicate strong dependence.

For large datasets (large n), we used l_b and u_b close together without incurring undue sampling variation in our statistics, and we considered multiple bands (l_b, u_b) to understand how dependence varies in different parts of the distributions. But for datasets with smaller n we considered only $l_b = 0$, $u_b = 0.5$ and $l_b = 0.5$, $u_b = 1$, abbreviating $\mathrm{cor}_l = \mathrm{cor}_{0,0.5}$ (l is for "lower") and $\mathrm{cor}_u = \mathrm{cor}_{0.5,1}$ (u is for "upper"). Likewise $\mathrm{P}_l = \mathrm{P}_{0,.5}$, $\mathrm{P}_u = \mathrm{P}_{0.5,1}$, $\mathrm{D}^2_l = \mathrm{D}^2_{0,0.5}$, $\mathrm{D}^2_u = \mathrm{D}^2_{0.5,1}$.

To test for asymmetry of dependence in upper and lower portions of distributions, we used differences $\mathrm{cor}_l - \mathrm{cor}_u$, $\mathrm{P}_l - \mathrm{P}_u$ and $\mathrm{D}^2_u - \mathrm{D}^2_l$ for smaller datasets (note the opposite order in the last of these); and $\mathrm{cor}_{0,u_b} - \mathrm{cor}_{1-u_b,1}$, $\mathrm{P}_{0,u_b} - \mathrm{P}_{1-u_b,1}$ and $\mathrm{D}^2_{1-u_b,1} - \mathrm{D}^2_{0,u_b}$ with u_b close to 0 for large datasets. We tested our statistics in Appendix S4 (see also Figs S23-S24). Code and data for this project is archived at www.github.com/sghosh89/BIVAN.

Results of testing the observational hypothesis: Ecological datasets often have non-normal copula structure and asymmetric tail dependence

Across essentially all our datasets, copula structures were typically non-N and commonly showed asymmetric tail dependence: the observational hypothesis (Fig. 2) was supported. To make results easier to absorb, we first present results for an

example bivariate dataset, then we summarize results for all bivariate datasets, then we present results for an example multivariate dataset, then we summarize for all multivariate datasets.

For the soil C and N data (Data, Table 1, Fig. 3A,E), variables were non-independent ($p = 0$, to within the precision available from `BiCopIndTest` in the `VineCopula` package); non-independence is also visually apparent. The Kendall correlation was 0.6. We fitted our 16 copulas to the data and computed AIC and BIC weights and corresponding lower- and upper-tail statistics for the fitted distributions (Table S3). BB1 was the minimum-AIC copula, with AIC weight 0.82, whereas the N copula had AIC weight essentially 0, confirming the first part of the observational hypothesis for these data. Model-averaged lower- and upper-tail dependence statistics were 0.453 and 0.615, respectively (AIC weights were used for averaging). These values are distinct from what an N copula would give, namely, 0, confirming the second part of the observational hypothesis. The values also differed substantially from each other, confirming the third part of the observational hypothesis. The numbers reflect stronger upper- than lower-tail dependence, and this is also visible in the extreme upper-right and lower-left corners of the copula plot (Fig. 3E).

Although these results convincingly show non-N copula structure, we caution that the model-averaged tail-dependence statistics may have been biased because even the lowest-AIC copula (BB1) was a poor fit ($p = 0$ and 0.001 for the Cramer-von Mises and Kolmogorov-Smirnov goodness-of-fit tests, respectively, to within the precision available from `BiCopGofTest`). However, nonparametric statistics verified both that tail dependence was distinct from that of an N copula, and that it was asymmetric (Appendices S5, S6, Fig. S25, Table S4).

Table 2 summarizes results of testing the observational hypothesis using the bivariate datasets. The two variables were significantly related for all datasets (Table 2, row 1). A non-N copula always emerged as the lowest-AIC (highest-AIC-weight) copula (Table 2, row 2). The N copula was always a poor fit compared to the highest-AIC-weight copula (Table 2, rows 3-4), confirming the first part of the observational hypothesis. Lower- and upper-tail dependence statistics often differed substantially from 0 (Table 2, rows 7-8) and from each other (Table 2, row 9), confirming the second and third parts of the observational hypothesis. The first three datasets had stronger upper- than lower-tail dependence, and the last dataset had the reverse, though weakly. Although these results convincingly show non-N copula structure, the model-averaged tail-dependence statistics may have been biased because even the lowest-AIC copula was a poor fit in some cases (Table 2, rows 5-6). But our nonparametric statistics confirmed the same conclusions (Appendices S5, S6). We did the same analysis with Cedar Creek data from other available years, 1996-1999 and 2007. For 1996 and 2007, biomass and Shannon's index were independent; for 1997-1999, results were similar to 2000: the best-fitting copula was not N, and there was slightly stronger lower- than upper-tail dependence. Detailed results for each dataset are in Tables S3 - S10 and Figs. S25 - S28.

We present green spruce aphid abundance (Table 1) results as an example multivariate analysis. Independence was rejected for each pair of the 10 sampling locations. Best-fitting (lowest-AIC) copulas were non-N for the large majority of location pairs (Table S11), and AIC weights for best-fitting copulas (Table S12) were typically substantially larger

than N-copula weights (Table S13), confirming the first part of the observational hypothesis. Goodness of fit tests (Tables S14, S15) in every case failed to reject the hypothesis that the AIC-best copula family was also an objectively adequate description of the data; the collection of copula families we used was sufficiently broad to characterize these data. Model-averaged lower- (Table S16) and upper-tail (Table S17) dependence statistics were exclusively positive and were frequently large, thereby differing from what an N copula would give (i.e., 0), and confirming the second part of the observational hypothesis. These data have greater tail dependence (lower and upper) than would be expected from an N-copula null hypothesis. Model-averaged lower- minus upper-tail dependence statistics (Table S18) were positive for all but a few location pairs (86 out of 90). Thus the spatial synchrony of rarity in the green spruce aphid is stronger than the spatial synchrony of outbreaks. This also confirms the third part of the observational hypothesis. Nonparametric statistics verified that tail dependence was asymmetric for the green spruce aphid abundance data (Appendix S7, S8).

Table 3 summarizes results of testing the observational hypothesis for multivariate data. Results supported the conclusions that non-N copula structure, non-N tail dependence, and asymmetric tail dependence were common, supporting the observational hypothesis. Most location pairs were non-independent (Table 3, row 2). The large majority of non-independent location pairs had best-fitting copulas that were not the N copula (Table 3, row 3), and AIC weights of best-fitting copulas were better than those of N copulas (Table 3, rows 4-5). Best-fitting copulas were nearly always considered an adequate fit (Table 3, row 6). Some of the datasets (green spruce aphid abundance, *Ceratium furca* abundance, methane-flux) showed stronger or marginally stronger lower- than upper-tail dependence (Table 3, rows 7-9), whereas leaf-curling plum aphid first flight data showed the reverse. Asymmetry results were verified by nonparametric approaches (Appendix S7, S8). Detailed results are in Tables S11-S46 and Figs S29-S32. We also carried out the same analyses for abundance and first-flight data for the 18 aphid species for which we had data other than the green spruce and leaf-curling plum aphids, as well as for the 21 plankton taxa for which we had abundance data other than the *Ceratium furca* (results not shown); conclusions supported that non-N copula structure and tail dependence and asymmetry of tail dependence are common.

Background, concepts and methods for testing the causal hypotheses

Having demonstrated that non-N copula structure and asymmetric tail dependence are common, we addressed the causal hypotheses (Fig. 2, A-D) using models. We did not systematically test the first two parts of A because nonlinear and thresold effects of environmental variables on ecological variables are already well documented (however, see Discussion material about the green spruce aphid). To test the third part of hypothesis A, we formulated a model in which an environmental variable influences an ecological variable more strongly in one of its tails, and explored whether the ecological variable could then exhibit asymmetric tail dependence across space. Let $E_i(t)$ be an ecological variable measured at location

i ($i = 1, 2$) at time t. Assume the dynamics $E_i(t+1) = bE_i(t) + g(\epsilon_i(t)) + a\delta_i(t)$, where the $\delta_i(t)$ are standard-normally distributed and independent across time and locations, $a = 0.2$, $b = 0.1, -0.1, 0.5, -0.5$ in different simulations, and the $(\epsilon_1(t), \epsilon_2(t))$ were drawn, independently through time, from a bivariate normal distribution with $\mathrm{var}(\epsilon_i) = 1$ and $\mathrm{cov}(\epsilon_1, \epsilon_2) = 0.8$. The function g describes how the environment $\epsilon_i(t)$ influences $E_i(t)$. We used g equal to g_1 or g_2 in different simulations: $g_1(\epsilon)$ equals ϵ if $\epsilon < 0$ and 0 otherwise; $g_2(\epsilon)$ equals ϵ if $\epsilon > 0$ and 0 otherwise. Thus g_1 represents environmental effects that occur only for negative values and g_2 only for positive values of ϵ. Negative values of b correspond to overcompensating dynamics and positive values to undercompensating dynamics; larger $|b|$ means slower return to equilibrium after a disturbance. Because (ϵ_1, ϵ_2) and (δ_1, δ_2) have N copula structure, the Moran mechanism analyzed below (Fig. 2, B) does not operate here.

For each b and g_i, we simulated the model for 25000 time steps and retained the $E_i(t)$ for the final 2500 time steps. We applied our model selection algorithms and non-parametric statistics (methods for the observational hypothesis) to these outputs to discover if the model could produce non-N copula structure and asymmetric tail dependence in the $E_i(t)$.

Hypothesis B is an extension of the well known Moran effect, and was summarized conceptually in the Introduction. Let $E_i(t)$ again be an ecological variable, $i = 1, 2$. We initially assume the AR(1) dynamics $E_i(t+1) = \beta E_i(t) + \sqrt{1-\beta^2}\epsilon_i(t)$, with $\beta = 0.5$. The environmental noises $\epsilon_i(t)$ are standard-normal random variables that are independent for distinct times, t, but exhibited different kinds of dependence across locations in different simulations (see below). The variable E is very general, and could represent deviations of a population density from a carrying capacity, deviations of total plant community biomass from an average value, flux of a biogeochemical variable such as methane, or other quantities.

To explore the importance of nonlinear dynamics for hypothesis B, we also considered nonlinear population models based on $P_i(t+1) = P_i(t)\exp\left(r\left(1 - P_i(t)/K\right) + \sigma\epsilon_i(t)\right)$, using $r = 0.5$, $K = 100$, and $\sigma = 0.1$ or $\sigma = 1$. When $\sigma = 0.1$, population dynamics stay close to the carrying capacity, K, and the nonlinearities of the model have limited influence on dynamics. When $\sigma = 1$, model dynamics are strongly nonlinear. We refer to these as the weak-noise and strong-noise cases, though the importance of the noise here is that, when it is strong, it brings the nonlinearities of the model into play.

For each of the model setups above, for each of several copula families, for each $\tau = 0.1, 0.2, \ldots, 0.9$, and for each of 50 replicate simulations, we generated 5000 noise pairs $(\epsilon_1(t), \epsilon_2(t))$ from the bivariate random variable with the given copula family and the given τ. We then used this noise to drive the model, and retained both the noise and population values for the final 500 time steps. For each simulation, the following statistics were computed for both noises and populations: Pearson, Spearman and Kendall correlations, cor_l, cor_u, P_l, P_u, D_l^2, D_u^2, $\mathrm{cor}_l - \mathrm{cor}_u$, $\mathrm{P}_l - \mathrm{P}_u$, and $\mathrm{D}_u^2 - \mathrm{D}_l^2$. Values were plotted against τ for noises and populations. If causal hypothesis B is correct, plots should be similar for populations and noises.

Hypothesis C was not tested. Our next causal hypothesis, Fig. 2, D, is evolutionary, and pertains to bivariate trait data across species, e.g., our bird and mammal data. This mechanism is a hypothetical explanation for the right tail dependence

observed in those data (Fig. 3F,G). The hypothesis is that tail dependence occurs in evolutionary changes in bivariate characters, and gives rise to tail dependence in observed character values across extant species. We simulated bivariate character evolution on an estimate of the phylogeny, taken from Genoud *et al.* (2018), of 817 mammal species. The root character state and the change across each branch were randomly chosen from matrices of one million bivariate draws from distributions showing one of five distinct types of copula structure: 1) extreme or 2) moderate left-tail dependence, 3) symmetric tail dependence, or 4) moderate or 5) extreme right-tail dependence (Appendix S9). All distributions had standard-normal marginals and Spearman correlation 0.875 between components, so our simulations assess the impact of copula structure only. For each of the five copulas, mammalian character evolution was simulated 100 times. For each simulation, symmetry of tail dependence of the two characters across phylogeny tips was assessed. We hypothesized that cases 1 and 2 above would yield stronger left- than right-tail dependence in tip characters, and cases 4 and 5 would yield the reverse. The simulator was written in Python and used version 4.4 of the DendroPy package (Sukumaran & Holder 2010).

Results of testing the causal hypotheses: Moran effects and asymmetric dependencies produce non-normal copula structure

Our model with asymmetric environmental effects could produce outputs with visually apparent asymmetry of tail dependence, to an extent that depended on the value of b, thus confirming the third part of hypothesis A for our model. For $b = 0.1$ (Fig. 4) and $b = -0.1$ (Fig. S33), for both g_1 and g_2, asymmetry of tail dependence was strong; for $b = \pm 0.5$, asymmetry was weaker but still apparent (Figs S34 - S35). Lower-tail (respectively, upper-tail) spatial dependence in effects of noise (g_1, respectively, g_2) produced lower-tail (respectively, upper-tail) spatial dependence in the ecological variable. Model selection results (Table S47) confirmed that ecological outputs had non-N copula structure, but did not always reflect asymmetry of tail dependence, perhaps because those methods measured dependence in the extreme tails. Results using our nonparametric statistics, however, reflected the visually apparent asymmetry (Tables S48).

Causal hypothesis B (Fig. 2) was largely confirmed; however, similarities between environmental- and ecological-variable copula structure were reduced when dynamics were strongly nonlinear. For simulations using AR(1) models, copula statistics were always very similar for noise and populations (Fig 5, S36-S41), though there were significant but minor differences for some statistics and simulations. For weak noise and our nonlinear model, copula statistics were again very similar for noise and model outputs (Figs S42- S53). Though there were again significant but small differences for some statistics and simulations, the hypothesis was supported that copula structure in the noise translates to similar but not identical copula structure in output populations. Since many ecological models are nonlinear, this result provides the expectation that there will be a Moran-effect-like correspondence between noise and model-output copula structure for

typical ecological dynamics, as long as noise is small enough that dynamics stay relatively close to the model equilibrium. Results that hold for "weak noise" in this sense are common in ecology. We repeated the analysis with $r = 1.3$, with similar results. The deterministic one-habitat-patch Ricker model exhibits a monotonic approach to a stable equilibrium at K when $r < 1$ (undercompensating dynamics, e.g., the value $r = 0.5$ used initially), but exhibits an oscillatory approach to K when $r > 1$ (overcompensating dynamics, e.g., $r = 1.3$).

For strong noise and using our nonlinear model, copula statistics, generally speaking, were approximately similar between noise and model outputs; however, similarity was reduced relative to previous results, and for a few simulations, tail dependence asymmetry statistics even had opposite signs for noise and model outputs (Figs S42-S53). For instance, using a C copula with a large Kendall τ, $\mathrm{cor}_l - \mathrm{cor}_u$ was slightly positive for noise, but slightly negative for population model outputs (Fig. S43); and likewise for SG and SJ copulas (Figs S49, S53). Overall, similarities between noise and model-output copula structure were more noticable than discrepancies, in spite of nonlinearities. We repeated the analysis with $r = 1.3$. Discrepancies between noise and population copula structure in this case were often glaring. Apparently noise of standard deviation 1 interacted strongly with model nonlinearities when the model was in its overcompensatory regime.

For our model, hypothesis D (Fig. 2) was correct that asymmetric tail dependence in evolutionary changes can produce similarly asymmetric tail dependence in character distributions across phylogeny tips (Appendix S9, Fig. S54 for details). This topic is revisited in the Discussion.

Background, concepts and methods for testing the consequential hypotheses

The hypothesis was presented that the distribution (through time) of a spatially averaged quantity should be influenced by dependencies between the local quantities, including their copula structure and tail dependence (Fig. 2, X). To refine this hypothesis, suppose an ecological variable $E_i(t)$ is measured at locations $i = 1, \ldots, N$ and times $t = 1, \ldots, T$, and the spatial mean $\sum_i E_i(t)/N$ is of interest. The $E_i(t)$ could be, for instance, local abundances of a pest or exploited species, or local fluxes of a greenhouse gas. If E_i and E_j are right-tail dependent for most location pairs i and j, then exceptionally large values tend to occur at the same time in most locations. We hypothesize that this will tend to increase the skewness of the distribution of the spatial mean. Similarly, left-tail dependence between local variables should tend to decrease skewness. Strong positive skewness of the spatial mean corresponds to "spikiness" of the spatial-mean time series. The spatial-mean time series and its skewness may be quantities of principal importance for pest or resource abundance, for which extreme values (spikes) in the spatial mean have large effects.

We tested hypothesis X using our multivariate datasets (Table 3). For each dataset we calculated the spatial mean time series, and then the skewness through time of that mean. Then, for each dataset, we compared the value obtained to a distribution of values of the same quantity for each of 10000 surrogate datasets. Surrogate datasets were produced by

randomizing the empirical data in a special way to have the copula structure of a multivariate-normal distribution, but to retain exactly the same distributions of values for each sampling location as the original data (Appendix S10). Surrogates also had very similar Spearman correlations between pairs of sampling locations as the data. Our comparisons therefore tested the null hypothesis that the skewness of the spatial mean took values on the empirical data no different from what one would expect if the copula structure of the data were multivariate N, but the data were otherwise statistically unchanged. Significant differences indicate that non-N copula structure in the data contributed to the skewness of the spatial mean, i.e., to its "spikiness".

For green spruce aphid abundance data, *C. furca* abundance data, and methane-flux data, because these datasets exhibited lower-tail dependence (Table 3), we compared empirical and surrogate skewness values via a one-tailed test in the left tail: the p-value was the fraction of surrogate skewnesses less than the skewness for the empirical data; the test examines whether lower-tail dependence caused the spatial average to have significantly lower skewness than would have been expected without tail dependence. For leaf-curling plum aphid data, because that dataset exhibited upper-tail dependence (Table 3), we did the analogous one-tailed test in the right tail. The test examines whether upper-tail dependence caused the spatial average to have significantly higher skewness than expected without tail dependence.

We also examined hypothesis Y (Fig. 2), i.e., that spatial tail dependence of an environmental variable can influence the extinction risk of a metapopulation. We hypothesized that environmental noises exhibiting greater left-tail dependence across different habitat patches would cause higher metapopulation extinction risks because then very bad years for the component populations would occur simultaneously in many patches, reducing rescue effects. Here we assume, for simplicity, that low values of the environmental variable are "bad" for the populations and high values are "good". We tested the reasonableness of Y using a metapopulation extension of the Lewontin-Cohen model, $\vec{P}(t+1) = D\lambda(t)\vec{P}(t)$, where the ith component of the length-N vector $\vec{P}(t)$ represents population density in the ith habitat patch at time t. The $N \times N$ matrix $\lambda(t)$ was diagonal with i^{th} diagonal entry $\exp(r + \epsilon_i(t))$. Here r is a growth rate; we used $r = 0$. The $\epsilon_i(t)$ represent environmental noises. They were standard-normally distributed, were independent through time, and showed the same spatial correlations for every simulation, but were made to exhibit right- or left-tail dependence between patches in different simulations (Appendix S11). The $N \times N$ matrix D was a dispersal matrix modelling local or global dispersal at rate d, in different simulations (Appendix S11). After each step, if the density in a patch was < 1, it was set to 0. We simulated the model 10000 times for each combination of parameters, starting from $p_0 = 50$ in each patch, and calculated extinction risk after 25 time steps.

Results of testing the consequential hypotheses: Tail dependence influences skewness of the spatial average, extinction risk, and Taylor's law

Skewness results were consistent with hypothesis X. For datasets that exhibited strong lower-tail (respectively, upper-tail) dependence in earlier analyses (green spruce aphid abundance data for lower-tail, and leaf-curling plum aphid first flight data for upper-tail, Table 3), skewness of the spatial average was less than (respectively, greater than) a significant fraction of surrogate skewness values (Fig. 6A,B). For datasets with moderate lower-tail dependence (*Ceratium furca* abundance and methane data), skewness of the spatial average showed a non-significant or marginally significant tendency toward being less than surrogate skewnesses (Fig. 6C,D).

Consistent with hypothesis Y, left-tail dependent environmental fluctuations increased metapopulation extinction risk for the spatial Lewontin-Cohen model (Fig. 7), for $N = 5$ and $N = 25$ habitat patches, and for local (Fig. 7) and global (Fig. S55) dispersal. We also substantiated using models (Appendix S12, S13, Fig. S56) a consequential hypothesis Z (Fig. 2), that Taylor's law can be influenced by tail dependence.

Discussion

We showed that non-N copula structure and asymmetric tail dependence are common across multiple sub-disciplines in ecology, although no research we know of has taken this structure into account. We hypothesized four mechanisms that may cause non-N copula structures. We discuss below how commonly some of our mechanisms may opperate. We also hypothesized and demonstrated important consequences of non-N copula structure and tail-dependence. For instance, the skewness of a spatial-average time series is influenced by tail dependencies between its constituent time series: right-tail-dependent local time series lead to "spiky" spatially averaged time series, with large outbreaks. Thus tail dependence could have implications for exploited and pest species. Extinction risk can be altered by tail dependence in environmental variables. In our view, our results make our overall hypothesis (Introduction) reasonable and worthy of broader testing. We illustrated how such tests might be carried out by performing tests ourselves on example datasets and models. Further suggestions for tests are below. It seems likely that comprehensive understanding of many ecological phenomena is impossible without a complete copula characterization of dependence.

Copula approaches are well developed (Nelsen 2006; Joe 2014), and computer implementations exist (`copula` and `VineCopula` packages). Ecologists can apply these tools immediately. We created several interrelated randomization procedures (Appendices S6, S10, S13) that build upon existing methods.

Our first causal mechanism (Fig. 2, A) probably operates commonly, for two reasons. First, Liebig's law and the idea of limiting nutrients are dominant paradigms in ecology. Second, although we did not systematically test the first two parts of hypothesis A, many studies have documented nonlinear or threshold influences of environmental variables on ecological

quantities. These first two parts of A could be tested for numerous datasets by examining copulas for environmental and ecological variables. See below discussion about the green spruce aphid, and Appendix S14, which illustrate how to do this. Causal mechanism C, which we also did not test, could be tested by analyzing copulas of abundances of competing species, sampled across space or time.

There are also reasons to expect our Moran mechanism (Fig. 2, B) operates commonly: Moran effects are common (Liebhold *et al.* 2004), and non-N and tail-dependent copula structures are often found in environmental variables. If intense meteorological events are also widespread, then environmental variables associated with these events should take extreme values simultaneously across large spatial areas, producing tail dependence in measurements made through time at different locations. Non-extreme values may instead be associated with local phenomena, and therefore may be less correlated across large areas. Serinaldi (2008) examined the spatial dependence of rainfall in Central Italy; G or Student 2-copulas were candidates for modelling dependence, and neither of these is an N copula. A long-term study (1950–2014) in the Loess Plateau of China (She & Xia 2018) showed that a G copula effectively modeled the spatial dependence of drought variables. Bivariate copula analysis was also used in forecasting the co-occurrence of extreme events (flood or drought) over the North Sikkim Himalayas using spatial datasets (Goswami *et al.* 2018).

Our simulations of character evolution suggest the hypothesis that changes through evolutionary time in bird and mammal BMR and body size may exhibit right-tail dependence, contrary to standard normality assumptions of character evolution models. This is a hypothesis only, because the right-tail dependence of Fig. 3F,G could have come about in another, unknown way. Our simulations show that tail dependence in evolutionary changes is sufficient, but may not be necessary, to produce tail dependence in characters of extant species. For instance (see below, Appendix S15, Fig. S57), systematically missing data can also produce tail dependence and may have influenced results for the BMR-body mass datasets. Even if the hypothesized evolutionary mechanism (Fig. 2, D) is correct, our results only replace one question, i.e., why do we see right-tail dependence between BMR and body mass, with another, i.e., why might we see right-tail dependence in evolutionary changes in these traits? See Appendix S9 for further thoughts on potential importance of copulas for character evolution.

Relationships between BMR and body mass relate to a trade-off between mass-specific BMR (BMR per unit body mass) and body mass itself. Copulas probably interrelate with life-history trade-off theory in additional ways beyond what we demonstrated. For instance, it is well known that energy allocation to a life function, F (e.g., reproduction) will reduce the energy that can be allocated to other functions, G_1, G_2, G_3 (e.g., growth, predation avoidance). This is the principle of allocation. But F can trade off against any or all of the G_i. Therefore, for large F, approaching absolute limitations, there may be a strong association between F and G_1. For small F, there may be little association because resources not allocated to F can instead be allocated to any combination of the G_i. This constitutes asymmetric tail dependence between F and G_1. Winemiller & Rose (1992) described a three-way trade-off in fishes between age of reproductive maturity, juvenile

survivorship, and fecundity. The trade-off should, in theory, produce a tight association between age of maturity and fecundity for fishes with low age of maturity, but little such association for later-maturing fishes because those fishes may invest the resources not invested in maturing quickly into either fecundity or juvenile survival. These ideas suggest that copulas will be useful for studying multi-dimensional trade-offs. But applications will require careful attention to the possible consequences of biased sampling: if the degree of completeness of a dataset is associated with one or more of the characters, then statistical artefacts can bias conclusions (Appendix S15, Fig. S57).

Other mechanisms probably also operate, and should be considered. For instance, measurement error may modify copula structure. Our models were intentionally simple in other respects, too, and did not include delayed density dependence, dispersal, population stage structure, trophic interactions, etc. We hope by enumerating a few potential mechanisms of copula structure, we inspire additional research on the potentially numerous mechanisms that may operate in diverse datasets.

Our hypotheses cover propagation of copula structure from causes, to data, to consequences. We here take a closer look at the green spruce aphid, which simultaneously illustrates all these stages. Green spruce aphid abundance, as measured in the data we use, is strongly influenced by temperature of the previous winter (Sheppard *et al.* (2016), supplementary fig. 6). For each of our 10 sampling locations, we therefore examined (Appendix S14, Fig. S58, Table S49) the copula of winter temperature and aphid abundance time series for the location, finding left-tail dependence: apparently winter temperature has an asymmetric influence on abundance in that cold winters produce low abundances but warm winters often do not yield higher abundances than moderate winters. Although we did not systematically test the first two parts of hypothesis A due to much existing evidence, this result illustrates how those parts can be tested using copula techniques. The third part of hypothesis A, which our modelling results supported, therefore suggests spatial dependence between green spruce aphid counts in different locations should show left-tail dependence - exactly as observed (Table 3). And the consequences of such tail dependence for the skewness of spatially averaged aphid counts was described previously (Fig. 6A). Thus the asymmetric influence of winter temperature ultimately causes spatially averaged aphid abundances to have lower skew than they would otherwise, because negative impacts of cold winters are spatially synchronous.

Our results showed that the skewness of the spatial average of local time series is influenced by their tail dependence. But the same logic should also apply to any collection of time series, whether associated with locations in space or not. Another potential application is time series of all species from a single community, e.g., all plants in a quadrat surveyed repeatedly over time. A huge literature has focused on synchrony versus compensatory dynamics between such time series, and the influence of interspecific relationships on the variability of community properties such as total biomass. Typically, variability of community biomass is measured with the coefficient of variation, but skewness may also be of interest because it can help characterize “spikiness” through time. Future work on copula structure of interspecific relationships in communities and its implications for community variability is likely to be valuable.

Although we demonstrated that tail dependence between environmental variables can influence extinction risk, substantial work remains to determine the importance of this effect. First, we used a non-density-dependent model. Do similar results pertain when density dependence acts? Second, we considered metapopulation extinction risk, but the huge field of population viability analysis (PVA) via stochastic matrix modelling (Caswell 2000; Morris & Doak 2002) uses a framework in which a single population's vital rates (e.g., life-stage-specific fecundity and survival rates) are considered to vary stochastically through time due to environmental variation. Do relationships between different vital rates exhibit tail dependence, and does tail dependence influence extinction risk in this context? Finally, is the copula structure of environmental variables changing through time, and, if so, how do such changes influence extinction risks? Climate change is known to amplify the factors that lead to extreme weather events (Hansen *et al.* 2012) and hence may alter tail dependence.

Acknowledgments

We thank the many contributors to the large datasets we used; D. Stevens and P. Verrier for data extraction; and Joel E. Cohen, David Tilman, Lauren Hallet, Jonathan Walter, Thomas Anderson, and Lei Zhao for helpful suggestions. We thank James Bell of the Rothamsted Insect Survey (RIS). The RIS, a UK Capability, is funded by the Biotechnology and Biological Sciences Research Council under the Core Capability Grant BBS/E/C/000J0200. SG, LWS and DCR were partly funded by US National Science Foundation grants 0075516 and 1442595 and the James S McDonnell Foundation.

References

Bell, J.R., Alderson, L., Izera, D., Kruger, T., Parker, S. & Pickup, J. *et al.* (2015). Long-term phenological trends, species accumulation rates, aphid traits and climate: Five decades of change in migrating aphids. *Journal of Animal Ecology*, 84, 21–34.

Burnham, K.P. & Anderson, D.R. (2003). *Model selection and multimodel inference: A practical information-theoretic approach.* Springer Science & Business Media.

Caswell, H. (2000). *Matrix population models: Construction, analysis, and interpretation.* Sinauer Associates, Sunderland, MA, USA.

Emura, T. & Chen, Y.-H. (2016). Gene selection for survival data under dependent censoring: A copula-based approach. *Statistical Methods in Medical Research*, 25, 2840–2857.

Genoud, M., Isler, K. & Martin, R.D. (2018). Comparative analyses of basal rate of metabolism in mammals: Data selection does matter. *Biological Reviews*, 93, 404–438.

Goswami, U.P., Hazra, B. & Goyal, M.K. (2018). Copula-based probabilistic characterization of precipitation extremes over North Sikkim Himalaya. *Atmospheric Research*, 212, 273–284.

Hansen, J., Sato, M. & Ruedy, R. (2012). Perception of climate change. *Proceedings of the National Academy of Sciences*, 109, E2415–E2423.

Harrington, R. (2014). The Rothamsted Insect Survey strikes gold. *Antenna*, 38, 159–166.

Holland, E.A., Robertson, G.P., Greenberg, J., Groffman, P., Boone, R. & Gosz, J. (1999). Soil CO_2, N_2O, and CH_4 exchange. In: *Standard soil methods for long-term ecological research* (eds. Robertson, G.P., Bledsoe, C.S., Coleman, D.C. & Sollins, P.). Oxford University Press, pp. 185–201.

Jarecke, K.M., Loecke, T.D. & Burgin, A.J. (2016). Coupled soil oxygen and greenhouse gas dynamics under variable hydrology. *Soil Biology and Biochemistry*, 95, 164–172.

Joe, H. (2014). *Dependence modeling with copulas.* Chapman & Hall/crc monographs on statistics & applied probability. Taylor & Francis.

Kim, J.-M., Jung, Y.-S., Sungur, E.A., Han, K.-H., Park, C. & Sohn, I. (2008). A copula method for modeling directional dependence of genes. *BMC Bioinformatics*, 9, 225.

Li, D. (2000). On default correlation: A copula function approach. *Journal of Fixed Income*, 9, 43–54.

Li, Y., Xie, K. & Hu, B. (2013). Copula-arma model for multivariate wind speed and its applications in reliability assessment of generating systems. *Journal of Electrical Engineering and Technology*, 8, 421–427.

Liebhold, A., Koenig, W.D. & Bjørnstad, O.N. (2004). Spatial synchrony in population dynamics. *Annual Review of*

Ecology, Evolution, and Systematics, 35, 467–490.

McNab, B.K. (2008). An analysis of the factors that influence the level and scaling of mammalian bmr. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, 151, 5–28.

McNab, B.K. (2009). Ecological factors affect the level and scaling of avian bmr. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, 152, 22–45.

Morris, W. & Doak, D. (2002). *Quantitative conservation biology: Theory and practice of population viability analysis*. Sinauer Associates, Sunderland, MA, USA.

Nelsen, R.B. (2006). *An introduction to copulas, second edition*. Springer.

Onken, A., Grünwälder, S., Munk, M. & Obermayer, K. (2009). Analyzing short-term noise dependencies of spike-counts in macaque prefrontal cortex using copulas and the flashlight transformation. *Plos Computational Biology*, 11, e1000577.

Serinaldi, F. (2008). Analysis of inter-gauge dependence by Kendall's τ K , upper tail dependence coefficient, and 2copulas with application to rainfall fields, 22, 671–688.

She, D. & Xia, J. (2018). Copulas-based drought characteristics analysis and risk assessment across the loess plateau of China. *Water Resources Management*, 32, 547–564.

Sheppard, L.W., Bell, J.R., Harrington, R. & Reuman, D.C. (2016). Changes in large-scale climate alter spatial synchrony of aphid pests. *Nature Climate Change*, 6, 610.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'institute de statistique de l'Université de Paris*, 8, 229–231.

Smyth, A.R., Loecke, T.D., Franz, T.E. & Burgin, A.J. (2019). Using high-frequency soil oxygen sensors to predict greenhouse gas emissions from wetlands. *Soil Biology and Biochemistry*, 128, 182–192.

Sukumaran, J. & Holder, M.T. (2010). DendroPy: a Python library for phylogenetic computing. *Bioinformatics*, 26, 1569–1571.

Tilman, D. (2018a). Plant aboveground biomass data: Biodiversity II: Effects of Plant Biodiversity on Population and Ecosystem Processes. Environmental Data Initiative. Available at : https://doi.org/10.6073/pasta/ab7b3604fb36e6875535fee68bd500a5.

Tilman, D. (2018b). Plant species percent cover data: Biodiversity II: Effects of Plant Biodiversity on Population and Ecosystem Processes. Environmental Data Initiative. Available at : https://doi.org/10.6073/pasta/ae57e6e9d87dd5fd204422eeec37b442.

Tilman, D. (2018c). Plot coordinates for the E120 Big Biodiversity Field: Biodiversity II: Effects of Plant Biodiversity on Population and Ecosystem Processes. Environmental Data Initiative. Available at :

https://doi.org/10.6073/pasta/940934edb437799bba2ca0a996cf072c.

Tilman, D., Reich, P.B. & Knops, J.M. (2006). Biodiversity and ecosystem stability in a decade-long grassland experiment. *Nature*, 441, 629.

Tilman, D., Reich, P.B., Knops, J., Wedin, D., Mielke, T. & Lehman, C. (2001). Diversity and productivity in a long-term grassland experiment. *Science*, 294, 843–845.

Walter, J., Sheppard, L., Anderson, T., Kastens, J., Bjørnstad, O. & Liebhold, A. *et al.* (2017). The geography of spatial synchrony. *Ecology Letters*, 20, 801–814.

Wills, S., Loecke, T., Sequeira, C., Teachman, G., Grunwald, S. & West, L.T. (2014). Overview of the U.S. Rapid Carbon Assessment Project: Sampling design, initial summary and uncertainty estimates. In: *Soil carbon* (eds. Hartemink, A.E. & McSweeney, K.). Springer International Publishing, Cham, pp. 95–104.

Winemiller, K. & Rose, K. (1992). Patterns of life-history diversification in North American fishes: Implications for population regulation. *Canadian Journal of Fisheries and Aquatic Sciences*, 49, 2196–2218.

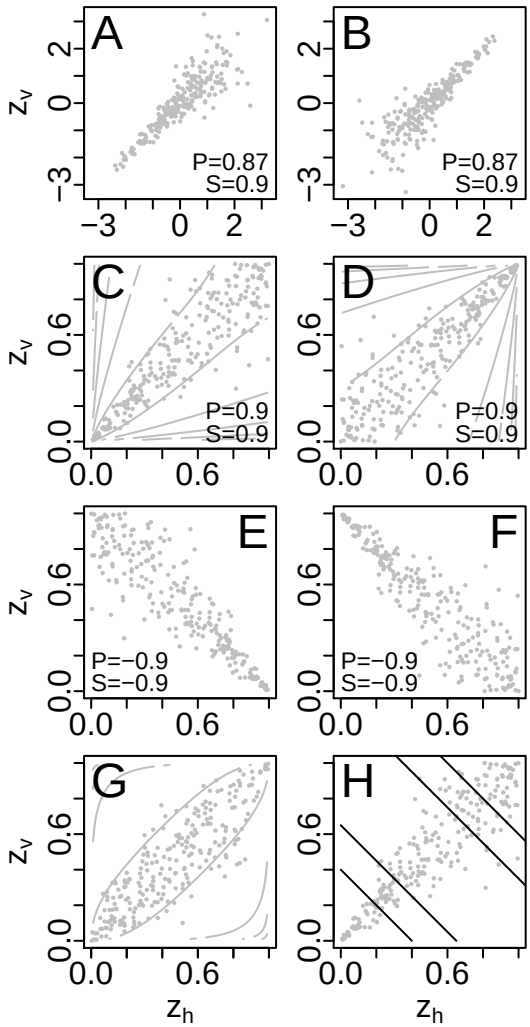

Figure 1: (A-D) Bivariate datasets showing diverse relationships between variables; many of these datasets nevertheless have the same Pearson (P) or Spearman (S) correlation coefficients. (C, D) Contours of probability density functions (pdf) of Clayton (C) and survival-Clayton (D) copulas with parameter 5 (see section on methods for the observational hypothesis), together with samples from the corresponding distributions. (E,F) Reflections of C, D, respectively, about a vertical axis. (G) Contours of the pdf of a normal copula (methods for the observational hypothesis) with parameter chosen to have the same Spearman correlation as A-D, together with a sample. (H) The portion of the Spearman correlation within a band can be computed for any band (methods for the observational hypothesis) to describe how dependence varies from one part of the two distributions to another. Diagonal lines show two bands, the data in the left/lower band showing stronger dependence than the right/upper band. Subscripts h and v stand for "horizontal" and "vertical".

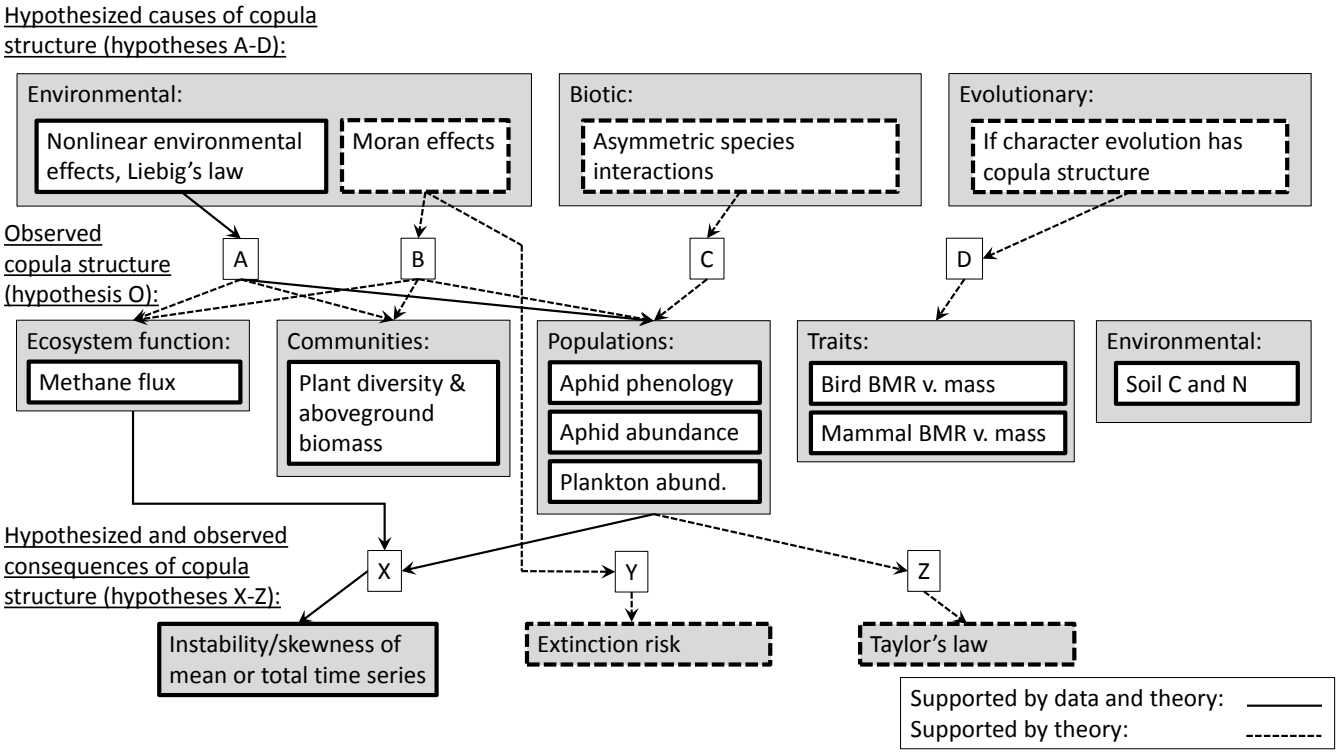

Figure 2: Guide for hypotheses addressed in the text and how each instantiates the overall hypothesis. Labels are defined in the text to refer to parts of the figure. "Causal hypotheses" (i.e., hypotheses about causes of copula structure) are labeled with letters A-D. "Consequential hypotheses" (i.e., hypotheses about consequences of copula structure) are labeled with X-Z. The "observational hypothesis" (i.e., our hypothesis about copula structure of data) is labeled O.

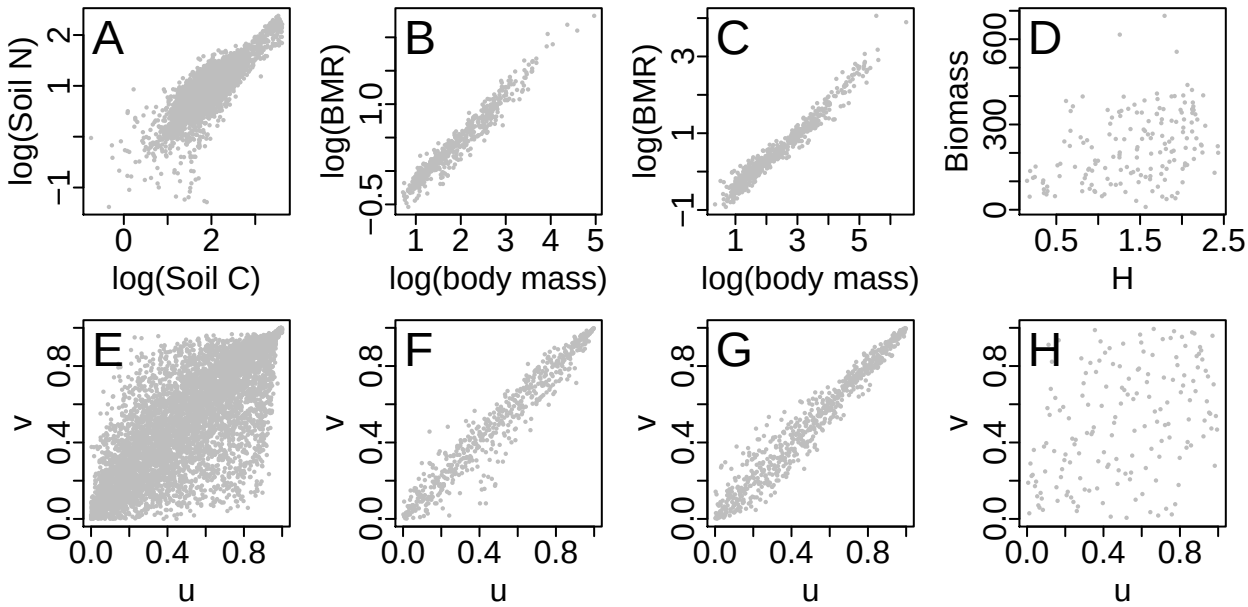

Figure 3: Upper panels show raw data plots for (A) $\log_{10}$(soil C) and $\log_{10}$(soil N) data, (B) $\log_{10}$(basal metabolic rate (BMR)) vs. $\log_{10}$(body mass) for birds, (C) the same for mammals, and (D) above-ground plant biomass vs. Shannon's index (H) from Cedar Creek. Bottom panels show the corresponding copula plots. See Data for units used in upper panels.

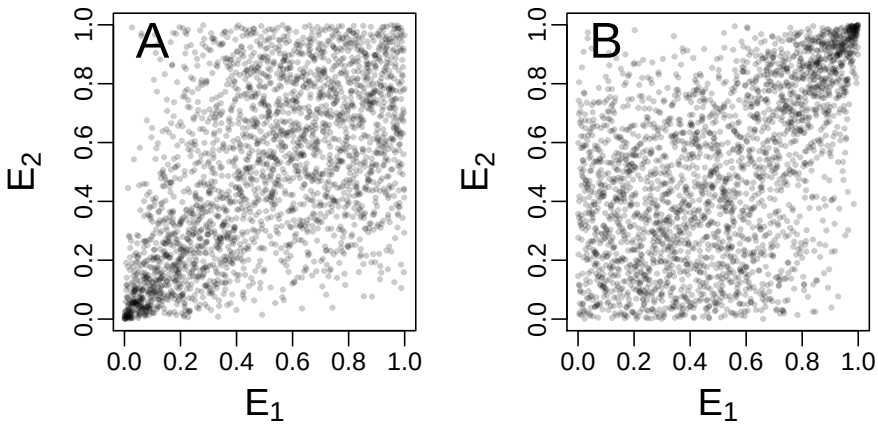

Figure 4: If environmental effects operate asymmetrically in their tails on ecological variables, it can result in non-N copula structure and asymmetric tail dependence across space in the ecological variables. Shown are the last 500 points for (A) $g = g_1$ and (B) $g = g_2$ from simulations described in the text, $b = 0.1$.

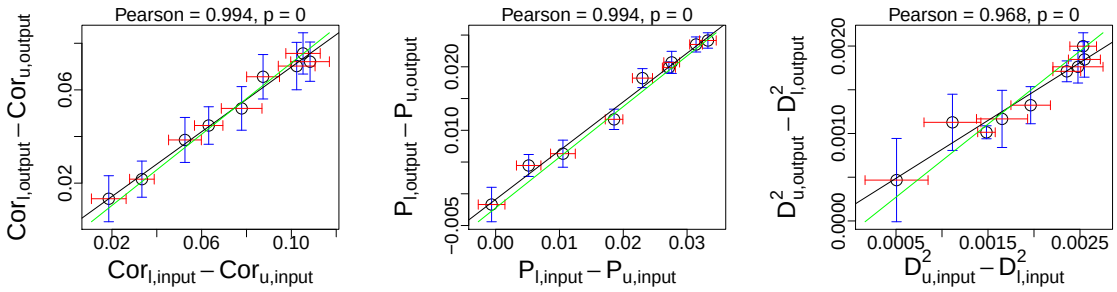

Figure 5: Example results showing similarity of copula statistics for environmental-noise inputs and ecological-variable outputs of a dynamical model. Asymmetric tail dependence was similar in both noise inputs and model outputs for this model. The AR(1) model and a Clayton copula were used (see methods section for the causal hypotheses). Each point is the mean across 50 replicate simulations for which the same Kendall τ value was used. Error bars are standard errors and panel headers give Pearson correlation results for the points. The regression line through the points (black line) was similar to the 1-1 line (green line). See Fig. S36 for additional results for the AR(1) model and Clayton copula, Figs S37-S41 for AR(1) results with other copula families, and Figs S42-S53 for the nonlinear models described in the methods for the causal hypotheses.

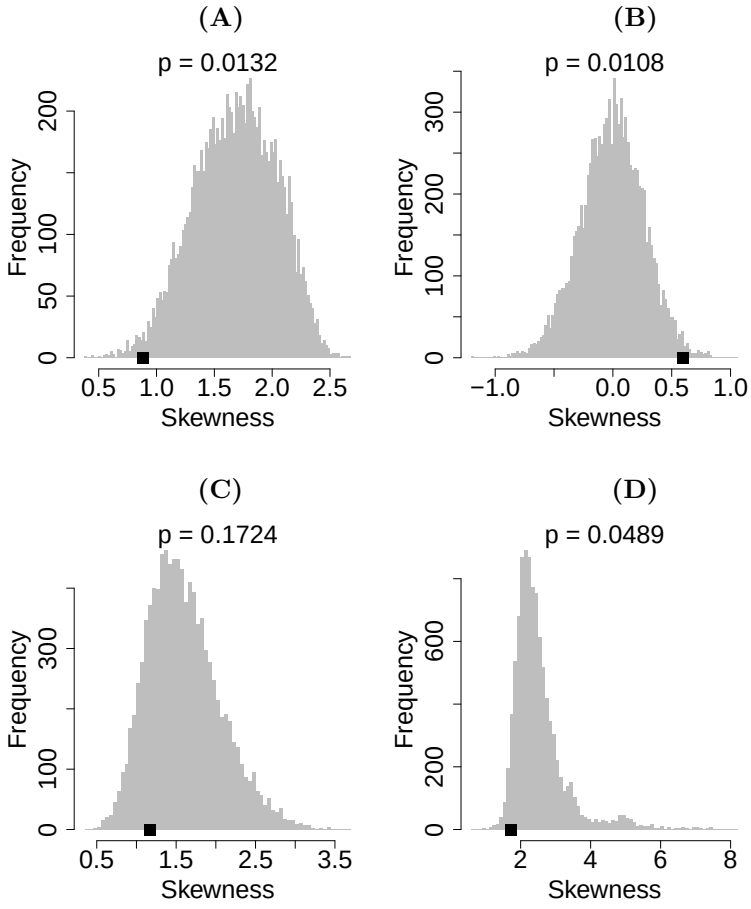

Figure 6: Skewness of spatially averaged green spruce aphid abundance (A), leaf-curling plum aphid first flight dates (B), *Ceratium furca* abundance (C), and methane-flux (D) time series compared to an multivariate N-copula null hypothesis. Black dots are empirical skewnesses; see text and Appendix S10 for details of the null hypothesis. Results show a tendency for skewness of the spatial average to be affected as hypothesized by tail dependence.

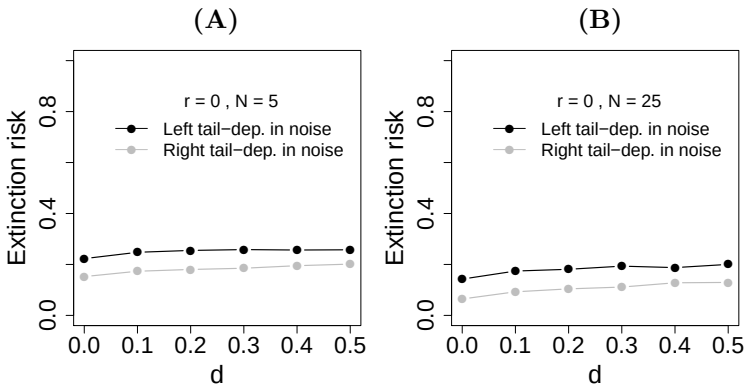

Figure 7: Extinction risk for the metapopulation extension of the Lewontin-Cohen model, after 25 time steps, was higher for left-tail-dependent environmental noise. Dispersal was local, with dispersal rate d (Appendix S11). Simulations used N patches for $N = 5$ (A) and $N = 25$ (B).

Table 1: Summary table for the data we used. Bold entries are the multivariate data, the rest are bivariate datasets (see Data section). Basal metabolic rate=BMR.

Data	Category	No. of measurements we used	References
Soil C and N	Environmental	5907 locations	US Rapid Carbon Assessment database (RaCA), Wills et al. (2014)
Bird body masses and BMR	Species-trait	533 birds	McNab (2009)
Mammal body masses and BMR	Species-trait	638 mammals	McNab (2008)
Green spruce and other aphid species abundances	**Population**	**10 locations with at least 30-yr. timeseries for each 20 sp.**	**Rothamsted Insect Survey**
Leaf-curling plum and other aphid first flight dates	**Phenological**	**10 locations with at least 30-yr. timeseries for each 20 sp.**	**Rothamsted Insect Survey**
***Ceratium furca* and other plankton taxa abundances**	**Population**	**14 locations with at least 45-yr. timeseries for each 22 taxa**	**Continuous Plankton Recorder Survey**
Plant diversity and aboveground biomass	Community level	168 plots	Cedar Creek Ecosystem Science Reserve, biodiversity experiment e120
Methane-flux	**Ecosystem functioning**	**13 locations with at least 50 dates of data**	**Great Miami Wetland Mitigation Bank, Smyth et al. (2019)**

Table 2: Summary of results of testing the observational hypothesis for bivariate datasets. The p-values in rows 5-6 are for the minimum-AIC copula. Model averaging used for rows 7-9 was based on AIC weights. Cedar Creek results are for the year 2000.

	Soil C and N	Bird masses and BMR	Mammal masses and BMR	Cedar Creek data
1. p-value, independence test	0	0	0	0
2. Minimum-AIC copula	BB1	G	G	F
3. AIC weight of minimum-AIC copula	0.816	0.578	0.597	0.378
4. AIC weight of N copula	0	0	0	0.122
5. p-value, Cramer-von Mises goodness of fit test	0	0.067	0.002	0.39
6. p-value, Kolmogorov-Smirnov goodness of fit test	0.001	0.279	0	0.67
7. Model averaged lower-tail dependence	0.453	0	0	0.055
8. Model averaged upper-tail dependence	0.615	0.877	0.893	0.003
9. Model averaged lower- minus upper-tail dependence	-0.162	-0.877	-0.893	0.052

Table 3: Summary of results of testing the observational hypothesis for multivariate datasets. Rows 3-9 of the table were computed for the non-independent pairs (row 2) only. A well-fitting location pair (row 6) was one for which the best-fitting copula had p-values > 0.01 for both the Cramer-von Mises and Kolmogorov-Smirnov goodness of fit tests.

	Green spruce aphid abundance	Leaf-curling plum aphid first flight	*Ceratium furca* abundance	Methane-flux
1. Location pairs (excluding self comparisons)	90	90	182	156
2. Number of non-independent pairs	90	90	162	102
3. Number of pairs with non-N copula as best fit	72	82	134	96
4. Average AIC-weights of best-fitting copulas across location pairs	0.215	0.232	0.221	0.266
5. Average AIC-weights of N copula across pairs	0.087	0.05	0.091	0.062
6. Number of well-fitting location pairs	90	83	160	102
7. 2.5^{th} and 97.5^{th} quantiles of model-avg. lower-tail dependence	(0.123 , 0.721)	(0 , 0.152)	(0.019 , 0.629)	(0.004 , 0.461)
8. 2.5^{th} and 97.5^{th} quantiles of model-avg. upper-tail dependence	(0.002 , 0.41)	(0.175 , 0.798)	(0 , 0.505)	(0 , 0.592)
9. Percent pairs showing stronger lower tail dependence	95.56	0	79.01	72.55