We thank all reviewers for their constructive comments on our manuscript. Below follows a point-by-point response to each comment.

Reviewer 1 (Theophile Caby)

General comments:
The manuscript presents a new methodological tool to compare compound extreme distributions between different datasets. The ability of a model to reproduce the behavior of compound extremes is of fundamental importance to assess climate related risks and to predict the evolution of such compound extreme events with climate change. The new metric is based on the Kullback-Leibler divergence. It is tested on different pairs of models and allows the comparison between different models regarding compound extreme distributions.

I find the manuscript well-motivated and clearly written, even for non-specialists of climate. The new metric seems promising and the statistical analysis made with it is well described and seems solid. The interpretation of the results is convincing to me, although my knowledge of climate models is limited.

Thank you.

Specific comments:
It is not mentioned whether the results are stable against different partitioning of the extremal region. You could add a few words about it: Are there partitions that are more suited than others? What made you chose this particular partition?

That is a very difficult theoretical question and the answer would depend somewhat on the extremal distribution of the two populations. As a rule of thumb, one would like to use as many sets as possible while guaranteeing that they still contain sufficiently many data points for a stable estimation of the probabilities that go into the Kullback-Leibler divergence. In response to RC3, we have conducted a small simulation study that revealed that $W \geq 5$ results in a robust test, and have therefore used $W = 5$ in the revision. This simulation study is now shown in the Appendix. The key results of the paper remain unchanged.

Technical corrections:
-l 29: 2 times the word 'studies'
-l 144: behavior
-l 240: may result

Thanks, we have incorporated the changes.

Reviewer 2

This manuscript compares the dependence structure of compound precipitation and wind speed extremes in different sets of data: the ERA5, the dynamically downscaled ERA-Interim using the regional WRF model, the dynamically downscaled CESM with present conditions using the WRF model, and also a dynamically downscaled CESM run for the future. The

technique used is an advanced statistical technique on bivariate asymptotic tail dependence. This is an interesting study which deserves publication in ESD. I have a few points on the interpretation of the results, and the limits of this study, that the authors may consider.

Thank you.

First, it seems to me strange to study extremes in a nudged system (ERAI-WRF). This means that there is a modification of the dynamical equations of WRF and the extremes could then be biased. First could you run it without the nudging, and if not this should appear somewhere in your interpretation or conclusions.

The reviewer is correct that the ERAI-WRF is nudged to the driving reanalysis ERA Interim. The reason for this is that the simulation should stay close to large-scale behavior of the reanalysis data. As mentioned in the manuscript, we only use wind, temperature and humidity above the planetary boundary layer and the nudging is not strong. So, we agree that to some extent the behavior of extremes might be changed due to the modification of the dynamical equations, but we think that this effect is minor. We also would like to point out that the precipitation is not nudged. To quantify the effect of nudging (and show that the effect is minor) a second simulation would be helpful, but currently we do not have the computational resources to perform such a simulation. We have added these explanations as penultimate paragraph in the discussion.

A second aspect is the fact that the domain over which downscaling is done looks small (no information provided on this by the way on the specific configuration of running WRF). This should have considerable impact on the extremes in particular for wind but also for precipitation. There were a lot of work done in this context at the beginning of the 21th century on that topic, showing that small domains are considerably constraining the internal dynamics of the regional model, and hence all the statistical properties within the model. This should also play an important role and should be discussed in the conclusions or in the interpretation of the results.

We see that the manuscript was not clear regarding the setup. We only show the innermost domain of a nested regional climate modelling approach using four nests in total. Domain 1 spans over Europe. The regions of the four nests are illustrated in the new Figure 1. In addition, we have added the following explanation: "The horizontal resolution of the four two-way nested domains (Fig. 1) are 54, 18, 6 and 2 km, respectively. The innermost domain 4 covers the box [4.75E,15.25E] × [43.25N,48.75N] and is used in this study exclusively."

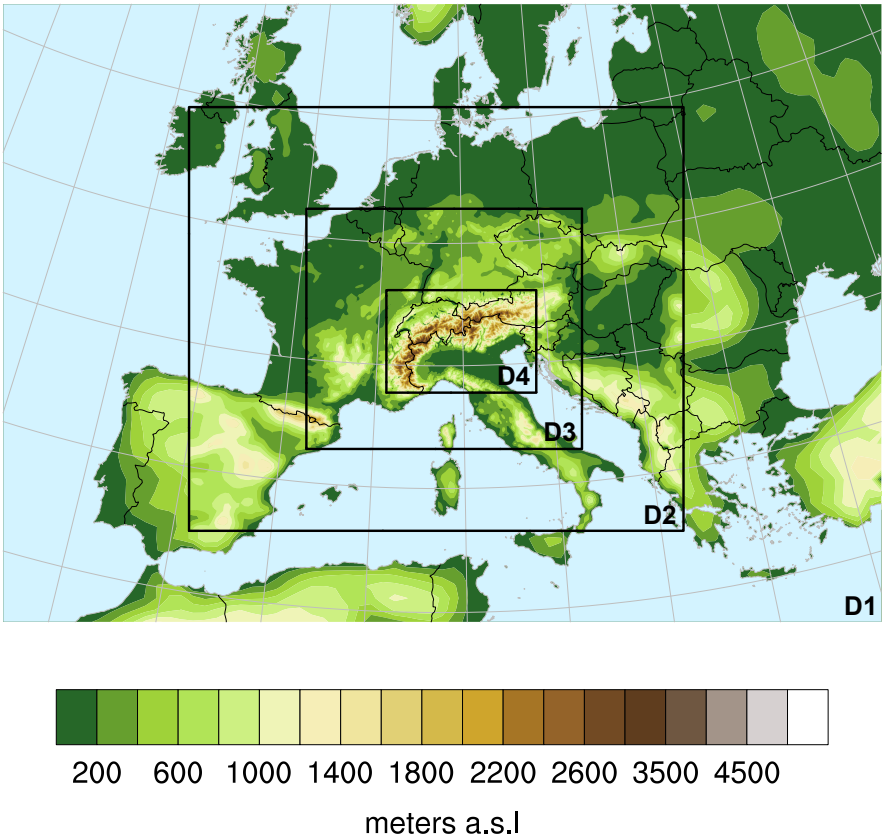


Figure 1: The four nested domains in used in the dynamical downscaling.

In Figure 1. It would be nice to see the observations too.

This is a pure modelling study and we do not have observational data for wind speed at different height at hand. This is why we refer to published results.

Short comment (Carlo de Michele)

The manuscript titled "Evaluating the dependence structure of compound precipitation and wind speed extremes" aims to estimate the likelihood of compound precipitation and wind speed extremes. In particular the Authors use metrics (the coefficients χ , $\bar{\chi}$, and KL measure of divergence) to measure the tail of bivariate distributions, and if it is similar between different datasets. The Authors use data from one reanalysis product (ERA5) and three model simulations (ERAI-WRF, CESM-WRF, CESM-WRF-fut) over a period of 20 years.

General comment: The manuscript is well written. The methodology is well grounded and of interest also for other compound events. The conclusions concerning the ERA5, which is considered as state-of-the-art dataset, are of value. Thus, I think that the manuscript should be accepted after minor revision.

Thank you.

Here you have some specific questions/comments:

- Application of the methodology to real data: Have the Authors an idea about the minimum number of data necessary to obtain a reliable estimate of the proposed metrics? Some details about this could be very useful.

This depends in general on the distribution of the data. For the χ coefficient we provide confidence intervals. For the KL statistics confidence interval can be obtained via bootstrap, and these will generally depend on the underlying distribution.

- Please give details about how you have calculated the KL measure of divergence, similar to the information given for the calculation of the coefficients χ and $\bar{\chi}$.

For computing the KL divergence, equation 1 and the one above provide the necessary equations to compute it.

- A reference for the risk function could be useful.

The sum is a classical way of looking at extremes, which treats all variables equally. If one projects all points that are large in this risk measure on the sphere r(x) = 1, then one obtains the so-called spectral measure, a popular object in extreme value theory. The minimum on the other hand excludes the axes and is therefore also suitable for asymptotically independent data.

- I think that in Figure 7 “K=3” should be substituted with “W=3”, to be coherent with the text. Similar comment applies to the caption of Table 1.

Thanks. Note that we use W=5 in the revision.

- Line 261 change “bivarate” with “bivariate”.

Thanks.

Reviewer 3

Disclaimer:
Even though I read the whole paper and appreciated both the methodological and applied aspects of this research, my review mostly revolves around the statistical contributions of this paper, which I’m more confident to comment on.

We highly appreciate the constructive comments.

Summary:
In this paper, the authors propose a new statistical metric to compare the bivariate joint tails of two different datasets. This metric, which relies on the Kullback-Leibler (KL) divergence based on the count of points in certain number of "extreme sets", provides a single number that can be used to assess whether or not the joint tails are different, and if so, by how much they differ. It is proposed as being complementary to more classical

measures, such as the χ-measure introduced in the paper that is widely used in extreme-value theory. The proposed KL metric depends on the number of sets, W, which has to be defined by the analyst, and is "non-parametric" in the sense that it does not rely on stringent model assumptions. In the paper, the proposed metric is used to estimate the likelihood of compound precipitation and wind speed extremes derived from different climate model outputs.

General assessment and general comments:
In my opinion, the paper is well written and concise with interesting practical results. Methodologically, the proposed metric is well-grounded but is not particularly novel as the Kullback-Leibler divergence (here based on the multinomial distribution) has been used extensively in other areas of statistics. The novelty probably relies on its specific use to study bivariate extremes and to compare bivariate joint tails of extreme precipitation and wind speed, although it is based on a previously published paper by one of the authors (Naveau et al., 2014, JRSS B). Overall, I like the paper and find the results quite interesting, yet several questions remain unanswered. My general and specific comments below mostly focus on the statistical contributions of the paper.

Thank you. Note that Naveau et al. 2014 only treated univariate times series, not bivariate (compound) events.

1. The χ measure is computed based on "local block maxima". I think it is easier to understand what the χ and $\bar{\chi}$ measures represent when used with the original daily data, rather than with block maxima. With original data, if χ = 0 this implies asymptotic independence of daily data, but how should we interpret it with block maxima? It would be good to add a few lines or a short paragraph to better explain the statistical meaning and the practical interpretation of the proposed metrics (χ, and KL-based) when they are used with block maxima. And why did you choose to compute χ based on block maxima and not block means or block minima? What is the rationale behind this choice? Is it somehow more informative to compare joint tails?

Taking block maxima is motivated by the underlying scientific question. We are interested in the relationship between (positive) extremes in precipitation and wind, which might not occur at the same time or at the same location but are driven by the same atmospheric process. These events can still cause disproportionate impacts. Furthermore, thresholding maxima implies that chi, chi-bar and KL really measure very extremal upper tail behavior. The drawback is the smaller sample size. However, this effect will also happen with thresholding, see e.g. Ferreira, A. and de Haan, L. (2015). On the block maxima method in extreme value theory: PWM estimators. The Annals of Statistics, 43(1):276–298.

Block maxima (instead of means or minima) are chosen because the interest is in the dependence between positive extremes of precipitation and wind speed.

We have added this motivation when introducing the block maxima approach in section 3.1 and section 3.2. In addition, we added the following note: "Note however, that this approach leads to different block sizes depending on the location, which makes a direct comparison in space difficult."

2. A major question that remains unclear to me is what do we gain with the proposed KL measure? As pointed out by the authors on page 5, we could compute a measure $\chi^{(1)}$ based on the first dataset, and another measure $\chi^{(2)}$ based on the second dataset and compare their values. The authors argue that they want just a single number to assess whether the tails are different and by how much. I get that. But why not simply doing a formal statistical test of whether $\chi^{(1)}$ is statistically different from $\chi^{(2)}$?? The test statistic (or the corresponding p-value) would indeed be a single value that could be used to assess whether the tails are different, and by how much. Moreover, the proposed KL metric is χ2-squared distributed ASYMPTOTICALLY, while testing for $\chi^{(1)} = \chi^{(2)}$ could—I think—be done EXACTLY for finite n (or be based on the corresponding asymptotic normal distribution). A partial answer to my question above ("what do we gain with the KL measure?") may be that the KL measure is probably more informative for testing whether the joint tails are different because it relies on full distribution of counts within extreme sets, rather than only on information about "the diagonal F1(X1) = F2(X2)"... but without a proper simulation study, this is difficult to claim (especially that the KL measure depends on the choice of W). It would be good if the authors could elaborate on that, and complement the paper with a short simulation study to assess the gains of the KL measure compared to a simple test $\chi^{(1)} = \chi^{(2)}$.

We appreciate the comment and suggestion. However, we believe that such a comprehensive simulation study would go beyond the interest of the readership of ESD. What one can say without a simulation study is that if we consider two distributions with the same χ coefficient but different dependence structure, then it is impossible to distinguish the two cases with the easier test the reviewer proposes. The referee is right when they state that χ focuses on the "diagonal". Furthermore, in this work we focus on the bivariate case, but the KL estimate defined by equation (1) could be easily implemented with higher dimensions d=3, 4, …, because it is just based on counting points in different subsets. With chi coefficients, the number of pairs will increase rapidly with d. In addition, chi coefficients will only capture pairwise dependencies. The KL does not have this problem and can easily be used for trivariate compounds events. We have added the following motivation for the KL divergence at the end of section 3.2:
"Note that in the bivariate case, a simple approach to quantify the difference in tail dependence would be the difference between $\chi^{(1)}$ and $\chi^{(2)}$. However, for two distributions with the same χ coefficient but different dependence structure, it is impossible to distinguish the two cases. In a way, χ only focuses on the 'diagonal'. Furthermore, while in this work we focus on the bivariate case, the KL divergence defined by (1) could be easily implemented with higher dimensions d=3, 4, …, because it is only based on counting points in different subsets. In contrast, using χ, the number of pairs will increase rapidly with the dimension d. In addition, χ coefficients will only capture pairwise dependencies."

3. This point is related to the point 2 above, but I split it into two parts so the authors can more easily address the several questions that I have. Another major question related to the proposed KL measure is how to set the number of extreme sets, W, to use. In the paper the authors choose W = 3, but there is no optimality with this choice. In fact, while the proposed KL measure is not well-defined when at least one of the sets is empty, the more classical χ-

measure is always well defined (so testing $\chi(1) = \chi(2)$ is always possible). This is a major drawback of the KL measure, I think, since under asymptotic independence we should EXPECT that the probability mass will concentrate on the axes (on the Pareto scale) with no point in the interior (so extreme sets should be empty in the limit!). Of course, in practice, there will always be points in the interior and ways to ensure that the extreme sets are non-empty, but it still raises the question of how to choose the number of sets W and the sets themselves. A related question is what is the efficiency of the statistical procedure for different numbers of sets, W ? In my opinion, it would be good to complement the paper with a simulation study, in order to investigate this issue in more details and come up with some concrete advice for practitioners on the selection of sets. Is there an "automatic" way to do this "well"?

At some level, χ is also based on an arbitrary choice because it is based on counting the number of points in the very specific "upper corner" ($X_1>u$, $X_2>u$)", given $X_1>u$. Our proposed KL divergence introduces more flexibility in terms of the choosing the norm, the number of set and the shape of sets. If the conditioning norm was equal to $r(x)=\min(x_1,0)$ and the partition just one set, $W = \{X_1>u, X_2>u\}$, then the KL measure will contain the same information than chi. Hence, instead of being a competitor, the KL measure broadens the scope of χ coefficients and allows for more detailed analysis. Of course, this added flexibility leads to more choices.

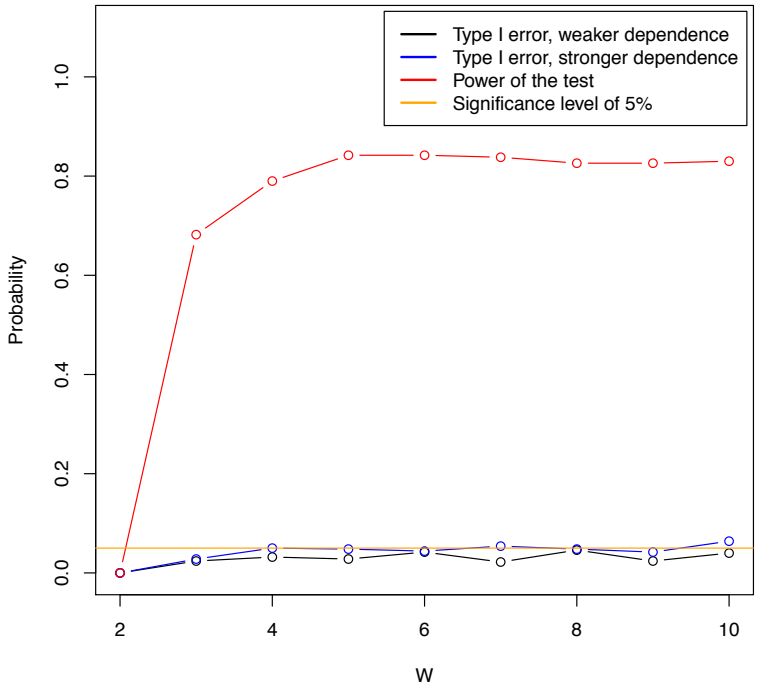
The case of asymptotic independence can be covered by our KL using $r(x) = \min(x_1, x_2)$ and by choosing sets W_i such that the probabilities of being no-empty in each set is positive. By assuming a second order type condition (classical multivariate EVT), Engelke, Naveau and Zhou (in prep) show that the convergence of our KL estimate towards a Chi-square distribution is still valid. For this theoretical statement, the marginals were supposed to be unknown with possibly different shape parameters. Hence, rank-based transforms were used, this answers the point 4 raised by the referee.
Under asymptotic dependence the empirical marginal normalization does have an effect on the asymptotic distribution. However, this effect is rather minor with little influence on the power of the test and the Type I error, as illustrated by a small simulation study as explained below.

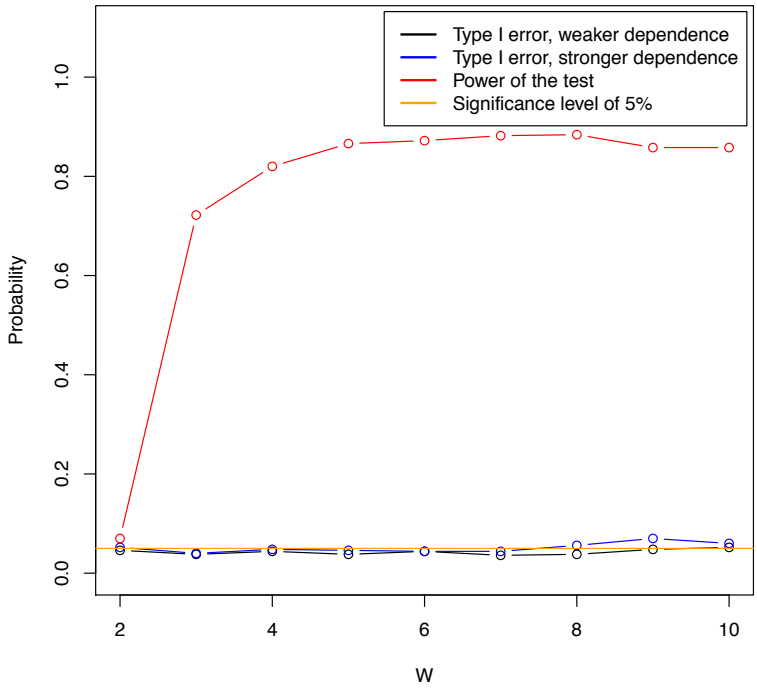
We simulated n=2000 samples $X^{(1)}$ and $X^{(2)}$ of the outer power Clayton copula, which is in the domain of attraction of the logistic extreme value distribution. We chose the parameters such that the limiting χ coefficients are 0.4 and 0.55, that is, one model with weaker and one with stronger dependence, respectively. Using the KL divergence for a probability threshold of u=0.9, we compare the samples of $X^{(1)}$ and $X^{(2)}$ for the dependence settings weak/weak, strong/strong and weak/strong and plot in each case the probability of rejecting the null hypothesis of equal tail dependence structures. Note that the former two cases are in line of the null hypothesis, whereas the latter case does not satisfy the null hypothesis. We do the experiment both for known margins and for empirically normalized margins, and for different numbers of sets W in the KL divergence statistic.

The two figures below show the Type I error of rejecting the null hypothesis in the case the where we have the same tail dependence based on 500 repetitions of the simulation. For both normalizations the significance level of 5% is in general well attained throughout all

numbers of sets. The figures also contain the power of the test when the tail dependence structures are different. After W=5 the power stabilizes and it seems to decrease slightly when the number of sets is chosen to large. We therefore suggest to take use W=5. Note that this is only one particular simulation setup and the results on the optimal number of sets can change depending on sample size and strength of tail dependence.

We have added these simulation results as appendix to the revised manuscript. Based on these simulation results, we now use W=5 in the revised version of the manuscript, which leads to a slightly higher number of significant KL divergences in Figure 8 and Table 1 but otherwise does not affect our main conclusions.





4. Another major point that is unclear to me is the treatment of marginal distributions. I assume that margins are estimated non-parametrically (with ranks) to compute the χ-measure, and that the extreme sets are defined based on data transformed to a common scale (e.g., Pareto), but there is no mention of marginal modeling in the paper. Does it matter here, since the KL-measure is non-parametric? I think this should be clarified. Marginal modeling usually has a major effect on the final results and their interpretation, so care is needed. In particular, how was the uncertainty related to marginal modeling taken into account (if it was)? The authors mention a bootstrap procedure for the χ-measure, but does it take marginal estimation uncertainty into account or does it only account for the estimation of the dependence structure?

The marginals have been transformed to Pareto scale through ranking. As stated in the response to 3., convergence of the KL estimates does not depend on this choice in the case of asymptotic independence. Under asymptotic dependence, empirical marginal normalization does change the asymptotic distribution but with only a very small effect on the robustness of the test, see the simulation study in response to comment 3, which has now been added as Appendix A to the manuscript. We have added the information how margins were transformed into Pareto scale at the end of section 3.2.

5. Figures 5-6: Even if I understand why the authors chose different block sizes (i.e., spatial lags and temporal windows), I find it difficult to interpret the results in Figure 5 given that the color in each pixel represents the tail dependence of potentially completely different events based on different block sizes. This may also explain why the figure looks a bit "noisy". Wouldn't it make more sense to produce such a map for each block size separately, and then present only the "most relevant" one (or potentially 2 block sizes of interest)? In my opinion, this would be much easier to interpret.

We agree that spatial points cannot be directly compared here as they might be based on different block sizes (as indicated in Figure 6). We believe however, that the "noisiness" is an actual signal, related to the extremely high resolution of the original data-generating process (2km) and the complex topography in the alps. This is supported by subpanel b), which is the only one based on much more coarse resolution data (~25km), and consequently shows much smoother spatial gradients (both in Figure 5 and 6). The choice of the block sizes is well motivated by the underlying scientific question (see response to comment 1 above).

6. Although the authors cite relevant papers related to extreme-value theory, some general review papers (or classical textbooks) could be added in my opinion to help non-experts navigate through this extensive literature.

We have added a general paragraph on extreme value theory (univariate and multivariate) at the beginning of section 3.1 referring to the following key literature:

Embrechts et al., 1997, Modelling Extremal Events: for Insurance and Finance (Springer)
Coles, 2001, An introduction to statistical modeling of extreme values (Springer)
Katz et al., 2002, Statistics of extremes in hydrology (AWR 25, 1287-1304)
Naveau et al., 2020, Statistical Methods for Extreme Event Attribution in Climate Science, (Annu. Rev. Statistics Appl., 7, 89–110)
Davison and Huser, 2015, Statistics of Extremes (Annu. Rev. Statistics Appl. 2, 203-235)
Huser and Wadsworth, in press, Advances in Statistical Modeling of Spatial Extremes (Interdisciplinary Reviews Computational Statistics)
Engelke and Ivanovs, 2021, Sparse Structures for Multivariate Extremes (Annu. Rev. Statistics Appl., in press)

Specific comments:
1. Page 2, Line 29: "studies studies"

Thanks.

2. Page 5, Line 119: If I'm not mistaken, the $\bar{\chi}$ measure has been introduced in a paper by Coles, Heffernan and Tawn (1999) published in Extremes, not by Ledford and Tawn (1996). Please add this reference.

Thanks, has been changed.

3. Page 5: Line 126 says "inspect their behavior as $q \rightarrow 1$" but Line 128 says "We generally estimate χ at $q = 0.95$". I agree and I get what the authors want to say, but these two sentences sound a bit contradictory. Please reformulate.

We reformulated the second sentence as "To estimate χ empirically we use a high quantile for which a reasonable large number of data are available. For these reasons we generally estimate χ at $q = 0.95$."

4. Page 5, Lines 149-150: you mention the sum and the minimum as the risk function r(x). Why not considering the maximum, as well, which is perhaps more commonly used than the minimum?

The sum or the maximum give similar results as they are both used for asymptotically dependent data. The minimum covers also asymptotic independence, and we have included it for this reason. We have added this explanation when introducing the risk function.

5. Page 6, Line 155: write "A_w(j)" instead of "A_w "?

Yes, thanks.

6. Page 6, Line 164, "The statistic d12 follows a $\chi 2(W - 1)$ distribution is the limit": Do you mean "in the limit as $n \rightarrow \infty$"? Also, is this valid under the null hypothesis that the tails are the same? Please clarify.

Yes, this is true under suitable assumptions, e.g., under asymptotic independence (with additional second order conditions) or if the data is multivariate regularly varying with the same marginal shape parameters (with additional second order conditions). Furthermore, $n \rightarrow \infty$ *and* $u(n) \rightarrow 1$ *need to converge at the right speed. We have added "Under suitable assumptions" to make clear that this convergence is conditioned to some general assumptions.*

7. Page 6, Lines 181-182, "q = 0.95 and u = 0.9": why did you choose different numbers? Does it matter?

These are somewhat arbitrary choices. We have carried out a sensitivity test for different values of u, which is shown in Table 1. Qualitatively the pictures doesn't change much (including its scientific interpretation) though of course the numbers are slightly different. In particular, with higher u, the number of significant KL divergences decreases, as is expected due to the smaller sample size. We have now added here that we conducted a sensitivity analysis with u in {0.8, 0.85, 0.9, 0.95}.

8. Page 7, Line 199: write "In particular, in the south of the Alps" (add "in the")

Thanks.

9. Page 7, Line 213-215: Table 1 shows the results are different as u increases. What do you conclude? And what if q increases?

The individual numbers change somewhat but the ranking within one column stays the same (except the flip of the first 2 rows at u=0.95, but both have a very similar value). The differences shown in row 1 and 3 are generally larger than the difference in row 4. This is the main scientific finding of the study, as also reported in the abstract: "Overall, boundary conditions in WRF appear to be the key factor in explaining differences in the dependence behaviour between strong wind and heavy rainfall between simulations. In comparison,

external forcings (RCP8.5) are of second order." We expect a very similar behavior for different values of q. We have added a sentence to make this finding more explicit: "In particular, the differences between ERAI-WRF and CESM-WRF and between ERAI-WRF and CESM-WRF-fut are generally larger than the differences CESM-WRF and CESM-WRF-fut, indicating that the main finding, namely that boundary conditions in WRF appear to be the key factor in explaining differences in the dependence behaviour between wind and rainfall extremes, is robust for different parameter values of the difference measure."

10. Page 8, Line 224: write "Because the model setting determines the dependence structure" (add "the")

Thanks.

11. Page 8, Lines 228-229: the sentence "This is to ensure ... (e.g., Foehn)" sounds odd to me. Please consider rewriting.

We have rewritten this sentence as "This is to ensure that extremes in wind and precipitation are considered together if they emerge from the same atmospheric processes (e.g. Foehn)."

12. Figure 3: The difference in tail behavior for the two datasets from q = 0.8 is already quite clear based on the χ-measure. This comes back to my general comments above: do we really need the new KL metric to detect this?

See our responses to the main comments above. Consider also the example where most of the data is above the diagonal in one case and below the diagonal in the other. Both distributions could have similar χ but the KL divergence would be large.

A marked-up manuscript version follows.

Evaluating the dependence structure of compound precipitation and wind speed extremes

Jakob Zscheischler[1,2,3], Philippe Naveau[4], Olivia Martius[1,5,6], Sebastian Engelke[7], and Christoph C. Raible[1,2]

[1]Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland
[2]Climate and Environmental Physics, University of Bern, Sidlerstrasse 5, 3012 Bern, Switzerland
[3]Department for Computational Hydrosystems, Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany
[4]Laboratoire des Sciences du Climat et de l'Environnement, Gif-sur-Yvette, France
[5]Institute of Geography, University of Bern, Bern, Switzerland
[6]Mobiliar Lab for Natural Risks, University of Bern, Bern, Switzerland
[7]Research Center for Statistics, University of Geneva, Geneva, Switzerland

Correspondence: Jakob Zscheischler (jakob.zscheischler@climate.unibe.ch)

Abstract. Estimating the likelihood of compound climate extremes such as concurrent drought and heatwaves or compound precipitation and wind speed extremes is important for assessing climate risks. Typically, simulations from climate models are used to assess future risks, but it is largely unknown how well the current generation of models represents compound extremes. Here, we introduce a new metric that measures whether the tails of bivariate distributions show a similar dependence structure across different datasets. We analyse compound precipitation and wind extremes in reanalysis data and different high-resolution simulations for central Europe. A state-of-the-art reanalysis dataset (ERA5) is compared to simulations with a weather model (WRF) either driven by observation-based boundary conditions or a global circulation model (CESM) under present-day and future conditions with strong greenhouse gas forcing (RCP8.5). Over the historical period, the high-resolution WRF simulations capture precipitation and wind extremes and ~~there~~ their response to orographic effects more realistically than ERA5. Thus, WRF simulations driven by observation-based boundary conditions are used as a benchmark for evaluating the dependence structure of wind and precipitation extremes. Overall, boundary conditions in WRF appear to be the key factor in explaining differences in the dependence behaviour between strong wind and heavy ~~rainfall~~ precipitation between simulations. In comparison, external forcings (RCP8.5) are of second order. Our approach offers new methodological tools to evaluate climate model simulations with respect to compound extremes.

1 Introduction

Compound extremes such as co-occurring drought and heat or compound precipitation and wind extremes can have substantial impact on the natural environment and human systems that often exceeds impact caused by a single extreme (Zscheischler et al.,

2014; Raveh-Rubin and Wernli, 2015; Martius et al., 2016; Sippel et al., 2018). Over the recent years a number of compound extremes have been investigated. For instance, several studies have analysed the dependence between storm surge and heavy precipitation (Wahl et al., 2015; Zheng et al., 2013; Bevacqua et al., 2019) or extreme runoff (Ward et al., 2018; Hendry et al., 2019) to estimate the risk of compound flooding in coastal areas. Compound droughts and heatwaves have been studied for different regions and varying temporal scales (Mazdiyasni and AghaKouchak, 2015; Zscheischler and Seneviratne, 2017; Manning et al., 2019; Sutanto et al., 2020; Zscheischler and Fischer, 2020). The occurrence rate of compound precipitation and wind extremes has been estimated for the Mediterranean region (Raveh-Rubin and Wernli, 2015), Europe (De Luca et al., 2020) and at the global scale (Martius et al., 2016). Other studies have investigated the co-occurrence of hot days and hot nights (Wang et al., 2020) or the co-occurrence rate of heavy precipitation and snow melt to estimate the risk of rain-on-snow events (Musselman et al., 2018; Poschlod et al., 2020). Such a quantification of the occurrence rate of compound extremes is important for assessing the risk of associated impacts today and in the future. Most of the above studies ~~studies~~ identify compound extremes by thresholding the contributing variables to quantify the occurrence of compound extremes and changes associated with climate change. However, the dependence structure in the tails only is rarely investigated. Due to the rarity of compound extremes, a large number of samples is required to obtain robust estimates, making it difficult to rely solely on observational data (Ridder et al., in press).

Large ensemble simulations (Deser et al., 2020) offer an opportunity to estimate future changes in the occurrence of compound events without running into data limitations (Poschlod et al., 2020; Champagne et al., 2020). However, such projections need to be interpreted with care as it is often largely unknown how well the employed models represent observed compound events (Musselman et al., 2018), and differences might be large between models. Climate models are regularly evaluated based on their ability to represent well-known processes in the climate system as well as predominantly univariate comparisons with key climate variables (Flato et al., 2013) though some multivariate metrics have been explored (Sippel et al., 2017). Yet little is known about the ability of climate models to capture observed occurrence rates of compound extremes (Zscheischler et al., 2018), a challenging task primarily for two reasons. First, due to their rarity, a robust quantification of the likelihood of compound extremes requires large amounts of data, thus making it difficult to establish "ground truth" for many applications. Second, suitable metrics for evaluating multivariate extremes have not been widely tested and applied in a climate context. Such metrics, however, are essential to assess how well models represent compound events, in particular to assess future risks (Zscheischler et al., 2020). When observational data are scarce, process-based model simulations (Couasnon et al., 2020) and reanalysis data (Martius et al., 2016) can be employed to extend or replace purely observational datasets.

To date, model-data comparisons related to compound extremes have been conducted to a very limited extent, often relying on simplifying assumptions and typically confined to precipitation and temperature. For instance, a high likelihood of compound hot and dry summers has been linked to a strongly negative correlation between summer temperature and precipitation (Zscheischler and Seneviratne, 2017). While there is generally a good agreement with respect to this correlation between climate models and observation-based datasets in the northern hemisphere, there is strong disagreement in the southern hemisphere, for which the models show a much stronger dependence. This finding may suggest that climate models overestimate dependence between summer temperature and precipitation. However, this discrepancy may also be related to the way gridded

observation-based datasets are assembled. In particular, for locations without an active measurement station nearby, the mean seasonal cycle is often used to fill gaps in the observational networks (e.g. Mitchell and Jones, 2005). This approach reduces the strength of co-variability between temperature and precipitation in poorly sampled regions, which are mostly in the southern hemisphere. Thus, assessing the ability of climate models to represent compound events may reveal underappreciated limitations in gridded observation-based datasets. We are not aware of studies so far that have evaluated the dependence between precipitation and wind speed.

In this study we focus on compound precipitation and wind extremes, which can have severe socio-economic impacts including human fatalities, impaired critical infrastructure and economical damage (Fink et al., 2009; Lin et al., 2010; Liberato, 2014; Raveh-Rubin and Wernli, 2015; Martius et al., 2016). We investigate differences in the occurrence of compound precipitation and wind extremes for different datasets over a region in central Europe around the Alps. To this end, we introduce a new measure that assesses dissimilarity between the tails of bivariate distributions. We study an experimental design with two factors. The first factor is the type of boundary conditions in a high-resolution regional weather model, either from reanalysis or a global circulation model. The second one corresponds to the effect of different climate forcing, between today and the future under a high-emission scenario. Our object of study under this design is the dependence between heavy rainfall and strong wind in winter over central Europe. In addition, comparisons with a state-of-the-art reanalysis product are implemented.

2 Data

We use daily precipitation sums and daily maximum wind speed in the extended winter (November-March) from one reanalysis product and three model simulations over a period of 20 years. The employed reanalysis product is the ERA5 data (Copernicus Climate Change Service (C3S), 2017) where we use the period 1980 to 1999 CE. This reanalysis is generated with an updated numerical weather prediction model and data assimilation system compared to the prior product ERA Interim (Dee et al., 2011) and integrates additional data sources. The data is available at resolution of roughly 30 km (spectral resolution of T639), 137 vertical levels and hourly output.

The three simulations are performed with the Weather Research and Forecasting (WRF) model (Skamarock et al., 2008) which is forced with boundary conditions from (i) ERA Interim (Dee et al., 2011) (ERAI-WRF), (ii) a period of free-running global climate simulation for present day (CESM-WRF) and (iii) a period covering the end of the 21st century under the Representative Concentration Pathway 8.5 (CESM-WRF-fut, a high-emission scenario). The global climate simulation is performed with the Community Earth System Model CESM (Hurrell et al., 2013) for the period 850 to 2100 CE. Details on the setting are described in Lehner et al. (2015) and Raible et al. (2018). In this study we use the periods 1980 to 1999 CE as present day and 2080 to 2099 CE as future.

The periods of the global simulations and the ERA Interim period (1980 to 1999 CE) are dynamically downscaled with WRF in version 3.5. WRF is vertically discretised in 40 terrain-following eta-coordinate levels. The horizontal resolution of the four two-way nested domains (Fig. 1) are 54, 18, 6 and 2 km, respectively. The innermost domain covers the box $[4.75°\mathrm{E}, 15.25°\mathrm{E}] \times [43.25°\mathrm{N}, 48.75°\mathrm{N}]$ and is used in this study exclusively. The setup is described in more detail in Gómez-

Navarro et al. (2015, 2018) and Messmer et al. (2017, 2020). Important for this a study is that the convection parameterisation is disabled for the simulations at 6 km and 2 km resolution; at these scales the model is convection-permitting. This is an important step in improving the simulation of precipitation, though still some problems remain (Ban et al., 2014). For simulating wind adequately, the setting of the planetary boundary layer parameterisation is key. We use a modified version of the fully non-local scheme developed by Hong and Lim (2020), which specifically treats effects of the unresolved orography (Jimenez and Dudhia, 2012). For the ERAI-WRF simulation we allow analysis nudging of wind, temperature and humidity above the planetary boundary layer, in order to stay close to large-scale behavior of the reanalysis data (Gómez-Navarro et al., 2015). For the two simulations driven by CESM, nudging is omitted to allow the regional model to correct potential systematic biases of the CESM (e.g., a too strong zonal atmospheric circulation in the mid latitudes (Bracegirdle et al., 2013)). The WRF output is provided in hourly resolution.

We remap the original hourly data to a common regular spaced grid with 0.25° spatial resolution using conservative remapping and subsequently compute daily precipitation sums and daily wind speed maxima. The 0.25° spatial resolution was chosen as it is closest to the original resolution of the ERA5 reanalysis data. Note however, that all WRF simulations are run on a much higher convection-resolving resolution. The explicit resolution of convection and a much higher resolution of the topography may result in a more accurate representation of the dependence between precipitation and wind extremes in the simulations than in ERA5. We further note that mean wind speed in ERA5 generally decreases with elevation (Fig. 2a), which is the opposite behaviour of what is the expected behaviour of the response of wind speed to elevation from observations (Graf et al., 2019; Telesca et al., 2020) and what is modelled by WRF (Fig. 2b). The discrepancy in mountainous regions between reanalysis data and observations with respect to wind speed is also evident in other reanalysis datasets such as ERA Interim (Jones et al., 2017), which is the predecessor of ERA5. In contrast, WRF has been shown to simulate wind speed reasonable well also in mountainous terrain (Stucki et al., 2016). For these reasons — WRF better resolves cloud processes, the topography and wind speed, ERA5 misrepresents wind speed gradient with elevation — we use ERAI-WRF as the reference for all analyses.

3 A measure for evaluating compound extremes

3.1 Measuring tail dependence

The extreme values of a univariate random variable can be analyzed with tools from extreme value theory (Embrechts et al., 1997; Coles, 20 . For multivariate random vectors, the dependence between the largest values in the components becomes important (Davison and Huser, 20 .

We quickly review the concept of bivariate asymptotic tail dependence and independence (Ledford and Tawn, 1997; Poon et al., 2003). Two variables X_1 and X_2 with cumulative distribution functions F_1 and F_2, respectively, are asymptotically dependent if

$$\chi = \lim_{q \to 1} \mathbb{P}(F_1(X_1) > q \mid F_2(X_2) > q) \ \in (0, 1],$$

and asymptotically independent otherwise (i.e., if $\chi = 0$). The coefficient χ is called extremal correlation and represents, after transforming X_1 and X_2 to the uniform scale, the probability of one variable being extreme given that the other one is extreme. Note that two variables can be dependent at normal levels but asymptotically independent in the extremes, as in the case for a bivariate Gaussian distribution (Sibuya, 1960). To fine tune the rate of decay towards the asymptotically independent case ($\chi = 0$), the residual tail dependence coefficient $\bar{\chi}$ contains additional information ~~(Ledford and Tawn, 1996)~~ (Coles et al., 1999):

$$\bar{\chi} = \lim_{q \to 1} \frac{\log(\mathbb{P}(F_1(X_1) > q)\mathbb{P}(F_2(X_2) > q))}{\log \mathbb{P}(F_1(X_1) > q, F_2(X_2) > q)} - 1 \ \in [-1, 1].$$

$\bar{\chi}$ is equal to 1 for asymptotically dependent variables, while for asymptotically independent variables $\bar{\chi}$ indicates if X_1 and X_2 are positively ($\bar{\chi} > 0$) or negatively ($\bar{\chi} < 0$) associated in their extremes. Thus, the pair of coefficients $(\chi, \bar{\chi})$ summarizes the tail dependence structure of X_1 and X_2.

Because both coefficients χ and $\bar{\chi}$ are defined as a limit value, a usual way to analyze the behaviour of a bivariate tail dependence structure between two variables is to compute empirical estimates for varying threshold levels q and then visually inspect their behaviour as $q \to 1$. We estimate χ and $\bar{\chi}$ with the function taildep from the R package extRemes (Gilleland and Katz, 2016).

~~We~~ To estimate χ empirically we use a high quantile for which a reasonable large number of data are available. For these reasons we generally estimate χ at $q = 0.95$. ~~To take into account that heavy~~ Heavy precipitation events and extreme winds that lead to large damages can be linked through storms or foehn events across neighboring locations and with a lag of several days. To take this aspect into account, we estimate χ using a local block maxima approach, which is motivated by (Ferreira and de Haan, 2015). We thus first compute the daily precipitation and wind speed maxima for varying block sizes ranging from 0.25° (approximately 20-30 km) to 1.75° (that is, maximum 3 grid points in any direction, or 100-200 km) and up to 5 days (i.e., maximum 2 days before and after the day of interest).

We further assess whether estimates of χ are significantly different from 0. To this end, we bootstrap the data by randomly shuffling the temporal order of one variable to break the dependence structure. The coefficient χ is then estimated as above. Estimates of χ are considered significantly different from 0 if they are larger than 95% of the bootstrapped estimates.

3.2 Measuring differences in bivariate extremal dependence structures

Classical tail coefficients like χ are informative summaries to assess the extremal dependence between two univariate random variables, say X_1 and X_2, but they cannot quantify the difference between extremal dependence between two **bivariate** random vectors, say $\boldsymbol{X}^{(1)} = (X_1^{(1)}, X_2^{(1)})$ and $\boldsymbol{X}^{(2)} = (X_1^{(2)}, X_2^{(2)})$. For example, a $\chi^{(1)}$ can be computed between heavy rainfall and strong winds computed from one dataset, e.g. ERA5, and compared to a $\chi^{(2)}$ for a second dataset, e.g. ERAI-WRF. But it would also be very convenient to have a single number to tell us if the extremal dependence between these two bivariate random vectors are different, and if so, by how much. Recent work by Naveau et al. (2014) showed the well-known Kullback–Leibler (KL) divergence used in signal processing can be tailored to the framework of extreme value theory. The approach has been applied to cluster climate data according to their bivariate extremal ~~bevaviour~~ behaviour (Vignotto et al., submitted).

However, to our knowledge, multivariate extremal divergence measures have never been applied to the analysis of compound weather and climate events. By complementing tail coefficients, this new tool could shed new lights on the joint behavior of heavy rainfall and strong winds across our different datasets.

The KL divergence is defined on marginals which are normalized to standard Pareto distributions. A risk function $r : \mathbb{R}^2 \to \mathbb{R}$ is used to describe the extreme region in each one of the bivariate distributions. ~~The risk functioncan be chosen as the sum $r(\boldsymbol{x}) = x_1 + x_2$ or the minimum~~ There are different choices for the risk function. Taking the sum or the maximum give similar results for asymptotically dependent data. In addition, the minimum covers also asymptotic independence. The sum is defined as $r(\boldsymbol{x}) = x_1 + x_2$, the minimum as $r(\boldsymbol{x}) = \min(x_1, x_2)$, $\boldsymbol{x} = (x_1, x_2)$. Hence, we consider as extreme points those for which the sum (or minimum) of the components exceeds a given high quantile $q_u^{(j)}$ of $r(\boldsymbol{X}^{(j)})$ corresponding to an exceedance probability $u \in (0,1)$, $j = 1, 2$. Varying the threshold $q_u^{(j)}$ alters the extremal region of interest. For each of the two bivariate distributions, the set $A^{(j)} = \{r(\boldsymbol{x}) > q_u^{(j)}\}$, $j = 1, 2$, is partitioned into a fixed number W of disjoint sets $A_1^{(j)}, \ldots, A_W^{(j)}$.

For two random samples $\boldsymbol{X}_1^{(1)}, \ldots, \boldsymbol{X}_n^{(1)}$, and $\boldsymbol{X}_1^{(2)}, \ldots, \boldsymbol{X}_n^{(2)}$, from the distributions $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$, the empirical proportions of data points belonging to set ~~A_W~~ $A_W^{(j)}$ is computed as

$$\widehat{p}_w^{(j)} = \frac{\#\left\{i : \boldsymbol{X}_i^{(j)} \in A_w^{(j)}\right\}}{\#\left\{i : r\left(\boldsymbol{X}_i^{(j)}\right) > q_u^{(j)}\right\}}, \quad w = 1, \ldots, W.$$

The difference between the extremal behaviours of the two distributions can then be measured as the KL divergence between the two multinomial distributions defined through these proportions, i.e.,

$$d_{12} = D(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}) = \frac{1}{2} \sum_{w=1}^{W} \left((\widehat{p}_w^{(1)} - \widehat{p}_w^{(2)}) \log(\widehat{p}_w^{(1)} / \widehat{p}_w^{(2)}) \right). \tag{1}$$

Note that this divergence is symmetric and since it is a non-parametric statistic it does not require additional model assumptions. Equation (1) contrasts differences among extremal dependence structures, both for asymptotically dependent and asymptotically independent data. The number of partitioning sets W is a free parameter. If it is chosen too high, many sets will be empty, resulting in an undefined KL divergence. If it is too small, only a rough summary is computed but not really an estimate of tail dependence. We chose ~~$W = 3$~~ $W = 5$ in this study ~~. The~~ based on the simulation study shown in Appendix A. Under suitable assumptions the statistic d_{12} follows a $\chi^2(W-1)$ distribution in the limit as the sample size goes to ∞, which allows us to estimate whether distances are significantly different from 0.

The approach is illustrated in Figs. 3 and 4. Figure 3 shows daily precipitation sums and maximum wind speed at grid point 9°E, 46.75°N on the original scale (a, d), and with margins normalized to exponential scale (b, e) and to standard Pareto distributions (c, f) for ERAI-WRF (a-c) and CESM-WRF (d-f). The shown grid point reaches the highest χ at $q = 0.95$ in the ERAI-WRF simulation. The colors in all subpanels and the dashed lines in Fig. 3c and f highlight the three disjoint sets $A_1^{(j)}, A_2^{(j)}$ and $A_3^{(j)}$, respectively (see above). At the exponential scale moderate and large extremes can be seen well whereas at the Pareto scale only very extreme values can be identified easily visually. Figure 4 illustrates χ (a) and $\bar{\chi}$ (c) for the distributions of the two simulations and the divergence based on Eq. (1) with "sum" (b) and "min" (d) as the risk function, including 95% confidence intervals of the empirical estimates. The estimates of χ and $\bar{\chi}$ start to diverge somewhat for $q > 0.8$, suggesting

different tail behavior (uncertainty ranges are estimated based on the R function chiplot from the package evd (Stephenson, 2002)). This impression is confirmed by the estimates of the KL divergence: For most thresholds $u > 0.5$ and both choices of the risk function the KL divergence is outside of the 95%-quantile of the limiting $\chi^2(W-1)$ distribution of the statistic d_{12} under the null hypothesis of equal tail dependence structures. This means that we can conclude that the two distributions have significantly different tail behaviour.

Note that in the bivariate case, a simple approach to quantify the difference in tail dependence would be the difference between $\chi^{(1)}$ and $\chi^{(2)}$. However, for two distributions with the same χ coefficient but different dependence structure, it is impossible to distinguish the two cases. In a way, χ only focuses on the "diagonal". Furthermore, while in this work we focus on the bivariate case, the KL divergence defined by Eq. (1) could be easily implemented with higher dimensions $d = 3, 4, \ldots$, because it is only based on counting points in different subsets. In contrast, using χ, the number of pairs will increase rapidly with the dimension d. In addition, χ coefficients will only capture pairwise dependencies.

We investigate how well different simulations represent the bivariate tail behaviour of daily precipitation sums and wind speed maxima in winter by comparing ERA5, CESM-WRF and CESM-WRF-fut against ERAI-WRF with the divergence as defined in Eq. (1) based the maxima over the spatio-temporal blocks that maximize tail dependence χ at $q = 0.95$. Using local block maxima ensures that χ, $\bar{\chi}$ and the KL divergence measure very extremal upper tail behavior. Note however, that this approach leads to different block sizes depending on the location, which makes a direct comparison in space difficult. For the computation of the KL divergence (Eq. (1)) we use $u = 0.9$ and "sum" as the risk function. We further perform a sensitivity test using $u \in \{0.8, 0.85, 0.9, 0.95\}$. Furthermore, the marginals have been transformed into Pareto scale through ranking. The choice of marginal transformation only has a minor influence on the KL divergence (see Appendix A).

4 Results

We first present a simple correlation analysis based on Spearman's rank correlation coefficient. Daily precipitation sums and maximum wind speed are generally strongly correlated in winter in most areas of the study domain except in the northwest of Italy (Fig. 5). All model simulations show a relatively consistent pattern, whereas ERA5 shows negative correlations at the southern slopes of the Alps along the northwestern Italian borders (Fig. 5b). Most correlations are significant ($\alpha = 0.05$).

When considering only the dependence in the tails based on χ and including a spatial and temporal neighborhood, the spatial patterns look quite different (Fig. 6). The WRF simulations show a highly heterogeneous picture with strong local variations, with generally strong dependence over most parts of the Alps and close to the Adriatic coast and weak dependence otherwise (Fig. 6a, c, d). Overall, ERAI-WRF shows slightly higher tail dependence compared to the WRF simulations driven by CESM. In contrast to the WRF simulations, in ERA5 tail dependence varies rather smoothly in space, with higher values in northeast Italy and along the eastern border of France (Fig. 6b).

The block sizes that attain the maximum tail dependence χ for precipitation and wind extremes for each pixel are shown in Fig. 7. On average for 75% of the pixels, the maximum is attained with no temporal lag. In contrast, there seems to be a shift in space, as maxima tend to co-occur in neighboring grid points: block sizes with larger than minimal (0.25°) spatial extent

occur on average in 60% of all locations (lighter colors in Fig. 7). This means that extremes in daily precipitation sums and wind extremes tend to occur on the same day but potentially not exactly at the same location but with some distance apart. In particular in the south of the Alps but also in some regions north of the Alps, this distance is 1.75°, or about 100-200 km (very light colors in Fig. 7). The strongest agreement of the dependence patterns exist between CESM-WRF and CESM-WRF-fut, which agree for half of the locations in the maximizing block size. In contrast, the agreement is 29% between ERAI-WRF and ERA5, and 39% between ERAI-WRF and CESM-WRF. Note that grid points at the boundaries cannot attain maxima with block sizes larger than one grid point as no data values are available outside the study domain.

The tails between winter daily precipitation sums and wind speed maxima show a significantly different dependence structure between ERAI-WRF and CESM-WRF in ~~40~~46% of all grid points, mostly in Switzerland and in the north of the study domain but also in many regions in northern Italy (Fig. 8a). The percentage of grid points with significantly different tail behaviour is slightly higher for the comparison of ERAI-WRF and ERA5 (~~42~~49%) though in this case most of the differences occur in grid points located along a wide diagonal band from south west to north east through the entire study domain (Fig. 8b). Interestingly, the comparison of ERAI-WRF with CESM-WRF-fut results in only ~~28~~36% pixels with significantly different tail behaviour (Fig. 8c). Thus, CESM-WRF-fut agrees better with ERAI-WRF with respect to the tail behaviour than CESM-WRF and ERAI-WRF. Finally, only ~~15~~18% of pixels show significantly different tail behaviour when comparing CESM-WRF and CESM-WRF-fut (Fig. 8d), indicating the pair with the largest number of grid points where no significant difference in the tail behavior could be found. The numbers of grid points with significantly different tail behaviour depends somewhat on the threshold u and generally decrease with increasing extremeness (that is, increasing u) but the differences between the different pairwise comparisons remains similar (Table 1). In particular, the differences between ERAI-WRF and CESM-WRF and between ERAI-WRF and CESM-WRF-fut are generally larger than the differences CESM-WRF and CESM-WRF-fut, indicating that the main finding, namely that boundary conditions in WRF appear to be the key factor in explaining differences in the dependence behaviour between wind and rainfall extremes, is robust for different parameter values of the difference measure.

5 Discussion

We have introduced a new metric for comparing tail dependence structures between wind and precipitation extremes in reanalysis data and weather model simulations. In our WRF simulations, the type of boundary conditions, either ERAI or CESM, appears to have a stronger effect on the coupling between high wind and heavy rainfall than the change of external forcing (present-day and future) in CESM (Fig. 8). This suggests that the studied dependence structures between the tails of precipitation sums and wind speed maxima in winter are a rather robust feature of the combination of models (boundary conditions plus high-resolution weather model) and thus also somewhat determined by the boundary conditions. In consequence this also means that here we are probably detecting rather stable dynamical features that are largely independent of strong external forcing such as (much) higher mean temperatures. Because the model setting determines the dependence structure, sampling

uncertainties in this dependence, for instance to robustly assess risks under future climate conditions, would require a range of different climate and weather model combinations.

The employed block maxima approach (Fig. 6) has the effect that precipitation and wind extremes are considered together even if they might occur some distance apart in either time or space. This is to ensure that ~~events~~ extremes in wind and precipitation are considered together ~~that likely~~ if they emerge from the same atmospheric processes (e.g. Foehn). At the same time, the block maxima approach can help diagnose why datasets differ in their tail dependence structure of precipitation and wind extremes, for instance if the spatio-temporal blocks for which extremes are attained differ strongly.

Regarding the optimal spatial and temporal lags between wind and precipitation extremes there is generally a good agreement that along the southern slopes of the Alps the dependence is maximized for precipitation and wind extremes occurring on the same day and up to 1.75° apart (lightest blue in Fig. 7), which could be related to Foehn events that lead to heavy precipitation north of the mountain range and extreme winds on the southern slopes or vice versa. Indeed, heavy precipitation events on the Alpine southside are often related to high moisture transport ahead of cold fronts that is associated with moderate winds that are not as strong as potential Foehn gusts on the Alpine north side (Panziera and Germann, 2010).

Most heavy precipitation events in the investigation domain in winter are associated with extratropical cyclones. Within extratropical cyclones, wind speed maxima and precipitation maxima are often linked to fronts and conveyor belts (Parton et al., 2010; Catto and Pfahl, 2013; Pfahl et al., 2014; Pantillon et al., 2020) and this may ~~results~~ result in co-located extremes. However, important modulations of both extreme wind and precipitation patterns by the local complex orography are to be expected (Whiteman, 2000; Barry, 2008) and such local Foehn effects, channelling effects, or flow blocking and many more might be captured by the high resolution WRF simulations but not in ERA5.

Overall, ERA5 shows quite a different behavior in Spearman's rank correlation (Fig. 5) and simple tail dependence χ (Fig. 6) compared to the high-resolution weather model simulations. Spatial patterns are much smoother, probably related to the much coarser spatial resolution (30 km compared to the original 2 km in the WRF simulations). Furthermore, wind speeds over high mountains are unrealistic, as they decrease with height rather than increase (Fig. 2). These limitations render ERA5 unsuitable as a benchmark for the tail dependence between precipitation and wind extremes in the Alpine area with its complex orography. Presently, homogenized gridded wind observations of good quality are not available for this region. Therefore, driving a well-calibrated high-resolution weather model with observation-based boundary conditions is currently the best benchmark to study compound wind and precipitation extremes.

We would like to note that in our setup ERAI-WRF is nudged to the driving reanalysis ERA Interim. The reason for this is that the simulation should stay close to large-scale behavior of the reanalysis data. As mentioned in the methods section, we only use wind, temperature and humidity above the planetary boundary layer and the nudging is not strong. Nevertheless, the behavior of extremes might be changed due to the modification of the dynamical equations to some extent, but we think that this effect is minor. Furthermore, precipitation is not nudged.

Evaluating how well models represent tail dependencies may help selecting those models that are fit for purpose (Maraun et al., 2017) regarding the analysis of compound events (Zscheischler et al., 2020) for a range of different event types (Ridder et al., in press). In particular, when the interest lies in the simulation of impacts, the approach may help decide when

multivariate bias adjustment approaches would need to be employed (François et al., 2020), as univariate bias adjustment might increase biases in impacts that depend on multiple correlated drivers (Zscheischler et al., 2019).

6 Conclusions

Evaluating the ability of climate models to represent the likelihood of compound climate extremes is important for well-informed climate risk assessments. In this study we investigated differences in the tail behaviour of precipitation and wind extremes in winter between different weather model simulations and a reanalysis dataset for a region in central Europe. Employing a new metric to measure differences in tail behaviour of ~~bivarate~~ bivariate distributions, we found that simulations of the same model pair with different external forcing conditions (climate change conditions) differ less than simulations for present-day conditions with different boundary data. Our results further suggest that reanalysis data are not suitable as a benchmark for the analysis of compound precipitation and wind extremes over complex terrain such as the Alps. Overall, differences between model simulations (different boundary conditions and weather/climate models) can be substantial. Our results suggest the climate impact modelling needs to take uncertainties related to the simulation of compound extremes into account to provide robust risk assessments for today and the future.

Data availability. ERA5 data are available from the ECMWF website: https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5. The output from the WRF simulations are very large data files and are available from Christoph Raible (christoph.raible@climate.unibe.ch).

Appendix A: Determining W

We simulated $n = 2000$ samples of $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ of the outer power Clayton copula, which is in the domain of attraction of the logistic extreme value distribution. We chose the parameters such that the limiting χ coefficients are 0.4 and 0.55, that is, one model with weaker and one with stronger dependence, respectively. Using the KL divergence for a probability threshold of $u = 0.9$, we compare the samples of $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ for the dependence settings weak/weak, strong/strong and weak/strong and plot in each case the probability of rejecting the null hypothesis of equal tail dependence structures. Note that the former two cases are in line with the null hypothesis, whereas the latter case does not satisfy the null hypothesis. We do the experiment both for known margins and for empirically normalized margins, and for different numbers of sets W in the KL divergence statistic.

Figure A1 and A2 show the Type I error of rejecting the null hypothesis in the case where we have the same tail dependence based on 500 repetitions of the simulation based on empirical ranking of the marginals and using the true marginals, respectively. For both normalizations the significance level of 5% is in general well attained throughout all numbers of sets. The figures also contain the power of the test when the tail dependence structures are different. After $W = 5$ the power stabilizes and it seems to decrease slightly when the number of sets is chosen to large. We therefore use $W = 5$ throughout the manuscript. Note that

this is only one particular simulation setup and the results on the optimal number of sets can change depending on sample size and strength of tail dependence.

Author contributions. J.Z. and P.N. conceived the idea and study design. P.N. and S.E. developed the code for the new metric. C.C.R. provided the model simulations. O.M. helped with the interpretation of the results. J.Z. performed all analysis, created all figures and wrote the first draft. All authors contributed substantially to the writing and revising of the manuscript.

Competing interests. The authors declare that they have no competing interests.

Acknowledgements. This research was supported by a Short-Term Scientific Mission from the European COST Action DAMOCLES (CA17109). We thank Martina Messmer for creating Figure 1. J.Z. acknowledges financial support from the Swiss National Science Foundation (Ambizione grant 179876). C.C.R is supported by the Swiss National Science foundation (grant: pleistoCEP – no. 200020_172745). The CESM and WRF simulations were performed on the supercomputing architecture of the Swiss National Supercomputing Centre (CSCS, Lugano, Switzerland. O.M. is supported by the Swiss National Science Foundation (grant 178751). Part of Philippe Naveau's research was supported by the FRAISE-LEFE-MANU grant and the french Agence National de la Recherche throughout the ANR-Melody and ANR-TRex.

References

Ban, N., Schmidli, J., and Schaer, C.: Evaluation of the convection-resolving regional climate modeling approach in decade-long simulations, Journal of Geophysical Research-Atmospheres, 119, 889–7907, https://doi.org/10.1002/2014JD021478, 2014.

Barry, R. G.: Mountain weather and climate, Cambridge University Press, Cambridge, 3rd edn., 2008.

Bevacqua, E., Maraun, D., Vousdoukas, M. I., Voukouvalas, E., Vrac, M., Mentaschi, L., and Widmann, M.: Higher probability of compound flooding from precipitation and storm surge in Europe under anthropogenic climate change, Science Advances, 5, eaaw5531, https://doi.org/doi/10.1126/sciadv.aaw5531, 2019.

Bracegirdle, T. J., Shuckburgh, E., Sallee, J.-B., Wang, Z., Meijers, A. J. S., Bruneau, N., Phillips, T., and Wilcox, L. J.: Assessment of surface winds over the Atlantic, Indian, and Pacific Ocean sectors of the Southern Ocean in CMIP5 models: historical bias, forcing response, and state dependence, Journal of Geophysical Research-Atmospheres, 118, 547–562, https://doi.org/10.1002/jgrd.50153, 2013.

Catto, J. L. and Pfahl, S.: The importance of fronts for extreme precipitation, Journal of Geophysical Research-Atmospheres, 118, 10 791–10 801, https://doi.org/10.1002/jgrd.50852, 2013.

Champagne, O., Leduc, M., Coulibaly, P., and Arain, M. A.: Winter hydrometeorological extreme events modulated by large-scale atmospheric circulation in southern Ontario, Earth System Dynamics, 11, 301–318, https://doi.org/10.5194/esd-11-301-2020, 2020.

Coles, S.: An introduction to statistical modeling of extreme values, Springer, https://doi.org/10.1007/978-1-4471-3675-0, 2001.

Coles, S., Heffernan, J., and Tawn, J.: Dependence measures for extreme value analyses, Extremes, 2, 339–365, 1999.

Copernicus Climate Change Service (C3S): ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate, Tech. rep., Copernicus Climate Change Service Climate Data Store (CDS), 2017.

Couasnon, A., Eilander, D., Muis, S., Veldkamp, T. I. E., Haigh, I. D., Wahl, T., Winsemius, H. C., and Ward, P. J.: Measuring compound flood potential from river discharge and storm surge extremes at the global scale, Natural Hazards and Earth System Sciences, 20, 489–504, https://doi.org/10.5194/nhess-20-489-2020, 2020.

Davison, A. and Huser, R.: Statistics of Extremes, Annual Review of Statistics and Its Application, 2, 203–235, 2015.

De Luca, P., Messori, G., Pons, F. M. E., and Faranda, D.: Dynamical systems theory sheds new light on compound climate extremes in Europe and Eastern North America, Quarterly Journal of the Royal Meteorological Society, https://doi.org/10.1002/qj.3757, 2020.

Dee, D. P., Uppala, S. M., Simmons, a. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. a., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, a. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, a. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, a. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Quarterly Journal of the Royal Meteorological Society, 137, 553–597, https://doi.org/10.1002/qj.828, 2011.

Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., Fiore, A., Frankignoul, C., Fyfe, J. C., Horton, D. E., Kay, J. E., Knutti, R., Lovenduski, N. S., Marotzke, J., McKinnon, K. A., Minobe, S., Randerson, J., Screen, J. A., Simpson, I. R., and Ting, M.: Insights from Earth system model initial-condition large ensembles and future prospects, Nature Climate Change, 10, 1–10, https://doi.org/10.1038/s41558-020-0731-2, 2020.

Embrechts, P., Klüppelberg, C., and Mikosch, T.: Modelling Extremal Events: for Insurance and Finance, Springer, London, 1997.

Engelke, S. and Ivanovs, J.: Sparse Structures for Multivariate Extremes, Annual Review of Statistics and Its Application, 8, in press.

Ferreira, A. and de Haan, L.: ON THE BLOCK MAXIMA METHOD IN EXTREME VALUE THEORY: PWM ESTIMATORS, The Annals of Statistics, 43, 276–298, http://www.jstor.org/stable/43556515, 2015.

Fink, A. H., Brücher, T., Ermert, V., Krüger, A., and Pinto, J. G.: The European storm Kyrill in January 2007: synoptic evolution, meteorological impacts and some considerations with respect to climate change, Natural Hazards and Earth System Sciences, 9, 405–423, https://doi.org/10.5194/nhess-9-405-2009, 2009.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., Forest, C., Gleckler, P., Guilyardi, E., Jakob, C., Kattsov, V., Reason, C., and Rummukainen, M.: Evaluation of Climate Models, in: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M., pp. 741–866, https://doi.org/10.1017/CBO9781107415324.020, 2013.

François, B., Vrac, M., Cannon, A. J., Robin, Y., and Allard, D.: Multivariate bias corrections of climate simulations: Which benefits for which losses?, Earth System Dynamics, 11, 537–562, https://doi.org/10.5194/esd-11-537-2020, 2020.

Gilleland, E. and Katz, R. W.: extRemes 2.0: An Extreme Value Analysis Package in R, Journal of Statistical Software, 72, 1–39, https://doi.org/10.18637/jss.v072.i08, 2016.

Gómez-Navarro, J. J., Raible, C. C., Bozhinova, D., Martius, O., García Valero, J. A., and Montávez, J. P.: A new region-aware bias-correction method for simulated precipitation in areas of complex orography, Geoscientific Model Development, 11, 2231–2247, https://doi.org/10.5194/gmd-11-2231-2018, 2018.

Gómez-Navarro, J. J., Raible, C. C., and Dierer, S.: Sensitivity of the WRF model to PBL parametrisations and nesting techniques: evaluation of wind storms over complex terrain, Geoscientific Model Development, 8, 3349–3363, https://doi.org/10.5194/gmd-8-3349-2015, 2015.

Graf, M., Scherrer, S. C., Schwierz, C., Begert, M., Martius, O., Raible, C. C., and Brönnimann, S.: Near-surface mean wind in Switzerland: Climatology, climate model evaluation and future scenarios, International Journal of Climatology, 39, 4798–4810, https://doi.org/10.1002/joc.6108, 2019.

Hendry, A., Haigh, I. D., Nicholls, R. J., Winter, H., Neal, R., Wahl, T., Joly-Laugel, A., and Darby, S. E.: Assessing the characteristics and drivers of compound flooding events around the UK coast, Hydrology and Earth System Sciences, 23, 3117–3139, https://doi.org/10.5194/hess-23-3117-2019, 2019.

Hong, S. and Lim, J.: The WRF single-moment 6-class micro-physics scheme (WSM6), Journal of Korean Meteorology Society, 42, 129–151, 2020.

Hurrell, J. W., Holland, M. M., Gent, P. R., Ghan, S., Kay, J. E., Kushner, P. J., Lamarque, J. F., Large, W. G., Lawrence, D., Lindsay, K., Lipscomb, W. H., Long, M. C., Mahowald, N., Marsh, D. R., Neale, R. B., Rasch, P., Vavrus, S., Vertenstein, M., Bader, D., Collins, W. D., Hack, J. J., Kiehl, J., and Marshall, S.: The Community Earth System Model A Framework for Collaborative Research, Bulletin of the American Meteorological Society, 94, 1339–1360, https://doi.org/10.1175/BAMS-D-12-00121.1, 2013.

Huser, R. and Wadsworth, J. L.: Advances in Statistical Modeling of Spatial Extremes, Interdisciplinary Reviews (WIREs) Computational Statistics, in press.

Jimenez, P. A. and Dudhia, J.: Improving the Representation of Resolved and Unresolved Topographic Effects on Surface Wind in the WRF Model, Journal of Applied Meteorology and Climatology, 51, 300–316, https://doi.org/10.1175/JAMC-D-11-084.1, 2012.

Jones, P. D., Harpham, C., Troccoli, A., Gschwind, B., Ranchin, T., Wald, L., Goodess, C. M., and Dorling, S.: Using ERA-Interim reanalysis for creating datasets of energy-relevant climate variables, Earth System Science Data, 9, 471–495, https://doi.org/10.5194/essd-9-471-2017, 2017.

Katz, R. W., Parlange, M. B., and Naveau, P.: Statistics of extremes in hydrology, Advances in Water Resources, 25, 1287–1304, 2002.

Ledford, A. W. and Tawn, J. A.: Statistics for near independence in multivariate extreme values, Biometrika, 83, 169–187, 1996.

Ledford, A. W. and Tawn, J. A.: Modelling dependence within joint tail regions, J. R. Stat. Soc. Ser. B Stat. Methodol., 59, 475–499, 1997.

Lehner, F., Joos, F., Raible, C. C., Mignot, J., Born, A., Keller, K. M., and Stocker, T. F.: Climate and carbon cycle dynamics in a CESM simulation from 850 to 2100 CE, Earth System Dynamics, 6, 411–434, https://doi.org/10.5194/esd-6-411-2015, 2015.

Liberato, M. L.: The 19 January 2013 windstorm over the North Atlantic: large-scale dynamics and impacts on Iberia, Weather and Climate Extremes, 5-6, 16 – 28, https://doi.org/https://doi.org/10.1016/j.wace.2014.06.002, 2014.

Lin, N., Emanuel, K. A., Smith, J. A., and Vanmarcke, E.: Risk assessment of hurricane storm surge for New York City, Journal of Geophysical Research: Atmospheres, 115, D18 121, https://doi.org/10.1029/2009JD013630, 2010.

Manning, C., Widmann, M., Bevacqua, E., Loon, A. F. V., Maraun, D., and Vrac, M.: Increased probability of compound long-duration dry and hot events in Europe during summer (1950–2013), Environmental Research Letters, 14, 094 006, https://doi.org/10.1088/1748-9326/ab23bf, 2019.

Maraun, D., Shepherd, T. G., Widmann, M., Zappa, G., Walton, D., Gutierrez, J. M., Hagemann, S., Richter, I., Soares, P. M. M., Hall, A., and Mearns, L. O.: Towards process-informed bias correction of climate change simulations, Nature Clim. Change, 7, 764–773, https://doi.org/10.1038/nclimate3418, 2017.

Martius, O., Pfahl, S., and Chevalier, C.: A global quantification of compound precipitation and wind extremes, Geophysical Research Letters, 43, 7709–7717, 2016.

Mazdiyasni, O. and AghaKouchak, A.: Substantial increase in concurrent droughts and heatwaves in the United States, Proceedings of the National Academy of Sciences, 112, 11 484–11 489, https://doi.org/10.1073/pnas.1422945112, 2015.

Messmer, M., Gómez-Navarro, J. J., and Raible, C. C.: Sensitivity experiments on the response of Vb cyclones to sea surface temperature and soil moisture changes, Earth System Dynamics, 8, 477–493, https://doi.org/10.5194/esd-8-477-2017, 2017.

Messmer, M., Gómez-Navarro, J. J., and Raible, C. C.: The Impact of Climate Change on the Climatology of Vb-Cyclones, Tellus, 14, in press, 2020.

Mitchell, T. D. and Jones, P. D.: An improved method of constructing a database of monthly climate observations and associated high-resolution grids, International Journal of Climatology, 25, 693–712, https://doi.org/10.1002/joc.1181, 2005.

Musselman, K., Lehner, F., Ikeda, K., Clark, M., Prein, A., Liu, C., Barlage, M., and Rasmussen, R.: Projected increases and shifts in rain-on-snow flood risk over western North America, Nature Climate Change, 8, 808–812, https://doi.org/10.1038/s41558-018-0236-4, 2018.

Naveau, P., Guillou, A., and Rietsch, T.: A non-parametric entropy-based approach to detect changes in climate extremes, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76, 861–884, 2014.

Naveau, P., Hannart, A., and Ribes, A.: Statistical Methods for Extreme Event Attribution in Climate Science, Annual Review of Statistics and Its Application, 7, 89–110, https://doi.org/10.1146/annurev-statistics-031219-041314, 2020.

Pantillon, F., Adler, B., Corsmeier, U., Knippertz, P., Wieser, A., and Hansen, A.: Formation of Wind Gusts in an Extratropical Cyclone in Light of Doppler Lidar Observations and Large-Eddy Simulations, Monthly Weather Review, 148, 353–375, https://doi.org/10.1175/MWR-D-19-0241.1, 2020.

Panziera, L. and Germann, U.: The relation between airflow and orographic precipitation on the southern side of the Alps as revealed by weather radar, Quarterly Journal of the Royal Meteorological Society, 136, 222–238, https://doi.org/10.1002/qj.544, 2010.

Parton, G., Dore, A., and Vaughan, G.: A climatology of mid-tropospheric mesoscale strong wind events as observed by the MST radar, Aberystwyth, Meteorological Applications, 17, 340–354, https://doi.org/10.1002/met.203, 2010.

Pfahl, S., Madonna, E., Boettcher, M., Joos, H., and Wernli, H.: Warm Conveyor Belts in the ERA-Interim Dataset (1979-2010). Part II: Moisture Origin and Relevance for Precipitation, Journal of Climate, 27, 27–40, https://doi.org/Doi 10.1175/Jcli-D-13-00223.1, 2014.

Poon, S.-H., Rockinger, M., and Tawn, J.: Extreme value dependence in financial markets: Diagnostics, models, and financial implications, The Review of Financial Studies, 17, 581–610, 2003.

Poschlod, B., Zscheischler, J., Sillmann, J., Wood, R. R., and Ludwig, R.: Climate change effects on hydrometeorological compound events over southern Norway, Weather and Climate Extremes, p. 100253, https://doi.org/https://doi.org/10.1016/j.wace.2020.100253, 2020.

Raible, C. C., Messmer, M., Lehner, F., Stocker, T. F., and Blender, R.: Extratropical cyclone statistics during the last millennium and the 21st century, Climate of the Past, 14, 1499–1514, https://doi.org/10.5194/cp-14-1499-2018, 2018.

Raveh-Rubin, S. and Wernli, H.: Large-scale wind and precipitation extremes in the Mediterranean: A climatological analysis for 1979-2012, Quarterly Journal of the Royal Meteorological Society, 141, 2404–2417, https://doi.org/10.1002/qj.2531, 2015.

Ridder, N., Pitman, A., Westra, S., Ukkola, A., Do, H., Bador, M., Hirsch, A., Evans, J., Luca, A. D., and Zscheischler, J.: Global hotspots for the occurrence of compound events, Nature Communications, in press.

Sibuya, M.: Bivariate extreme statistics, Annals of the Institute of Statistical Mathematics, 11, 195–210, 1960.

Sippel, S., Zscheischler, J., Mahecha, M. D., Orth, R., Reichstein, M., Vogel, M., and Seneviratne, S. I.: Refining multi-model projections of temperature extremes by evaluation against land–atmosphere coupling diagnostics, Earth System Dynamics, 8, 387–403, https://doi.org/10.5194/esd-8-387-2017, 2017.

Sippel, S., Reichstein, M., Ma, X., Mahecha, M. D., Lange, H., Flach, M., and Frank, D.: Drought, Heat, and the Carbon Cycle: a Review, Current Climate Change Reports, 4, 266–286, https://doi.org/10.1007/s40641-018-0103-4, 2018.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Wang, W., and Powers, J. G.: A description of the advanced research WRF version 3, Tech. rep., TN-475+STR, National Center for Atmospheric Research, 2008.

Stephenson, A. G.: evd: Extreme Value Distributions, R News, 2, 0, https://CRAN.R-project.org/doc/Rnews/, 2002.

Stucki, P., Dierer, S., Welker, C., Gómez-Navarro, J. J., Raible, C. C., Martius, O., and Brönnimann, S.: Evaluation of downscaled wind speeds and parameterised gusts for recent and historical windstorms in Switzerland, Tellus A: Dynamic Meteorology and Oceanography, 68, 31 820, https://doi.org/10.3402/tellusa.v68.31820, 2016.

Sutanto, S. J., Vitolo, C., Napoli, C. D., D'Andrea, M., and Lanen, H. A. V.: Heatwaves, droughts, and fires: Exploring compound and cascading dry hazards at the pan-European scale, Environment International, 134, 105 276, https://doi.org/https://doi.org/10.1016/j.envint.2019.105276, 2020.

Telesca, L., Guignard, F., Laib, M., and Kanevski, M.: Analysis of temporal properties of extremes of wind measurements from 132 stations over Switzerland, Renewable Energy, 145, 1091 – 1103, https://doi.org/https://doi.org/10.1016/j.renene.2019.06.089, 2020.

Vignotto, E., Engelke, S., and Zscheischler, J.: Clustering bivariate dependences in the extremes of climate variables, Journal of Climate, submitted.

Wahl, T., Jain, S., Bender, J., Meyers, S. D., and Luther, M. E.: Increasing risk of compound flooding from storm surge and rainfall for major US cities, Nature Climate Change, 5, 1–6, https://doi.org/10.1038/nclimate2736, 2015.

Wang, J., Chen, Y., Tett, S. F., Yan, Z., Zhai, P., Feng, J., and Xia, J.: Anthropogenically-driven increases in the risks of summertime compound hot extremes, Nature Communications, 11, https://doi.org/10.1038/s41467-019-14233-8, 2020.

Ward, P. J., Couasnon, A., Eilander, D., Haigh, I. D., Hendry, A., Muis, S., Veldkamp, T. I., Winsemius, H. C., and Wahl, T.: Dependence between high sea-level and high river discharge increases flood hazard in global deltas and estuaries, Environmental Research Letters, 13, 084 012, 2018.

Whiteman, C. D.: Mountain meteorology fundamentals and applications, Oxford University Press, New York, 2000.

Zheng, F., Westra, S., and Sisson, S. A.: Quantifying the dependence between extreme rainfall and storm surge in the coastal zone, Journal of hydrology, 505, 172–187, 2013.

Zscheischler, J. and Fischer, E.: The record-breaking compound hot and dry 2018 growing season in Germany, Weather and Climate Extremes, 19, 100 270, https://doi.org/10.1007/s00484-020-01951-8, 2020.

Zscheischler, J. and Seneviratne, S. I.: Dependence of drivers affects risks associated with compound events, Science Advances, 3, e1700 263, 2017.

Zscheischler, J., Michalak, A. M., Schwalm, C., Mahecha, M. D., Huntzinger, D. N., Reichstein, M., Berthier, G., Ciais, P., Cook, R. B., El-Masri, B., Huang, M., Ito, A., Jain, A., King, A., Lei, H., Lu, C., Mao, J., Peng, S., Poulter, B., Ricciuto, D., Shi, X., Tao, B., Tian, H., Viovy, N., Wang, W., Wei, Y., Yang, J., and Zeng, N.: Impact of large-scale climate extremes on biospheric carbon fluxes: An intercomparison based on MsTMIP data, Global Biogeochemical Cycles, 28, 585–600, https://doi.org/10.1002/2014GB004826, 2014.

Zscheischler, J., Westra, S., Hurk, B. J., Seneviratne, S. I., Ward, P. J., Pitman, A., AghaKouchak, A., Bresch, D. N., Leonard, M., Wahl, T., and Zhang, X.: Future climate risk from compound events, Nature Climate Change, 8, 469–477, 2018.

Zscheischler, J., Fischer, E. M., and Lange, S.: The effect of univariate bias adjustment on multivariate hazard estimates, Earth System Dynamics, 10, 31–43, 2019.

Zscheischler, J., Martius, O., Westra, S., Bevacqua, E., Raymond, C., Horton, R. M., van den Hurk, B., AghaKouchak, A., Jézéquel, A., Mahecha, M. D., Maraun, D., Ramos, A. M., Ridder, N., Thiery, W., and Vignotto, E.: A typology of compound weather and climate events, Nature Reviews Earth & Environment, 1, 333–347, https://doi.org/10.1038/ s43017-020-0060-z, 2020.

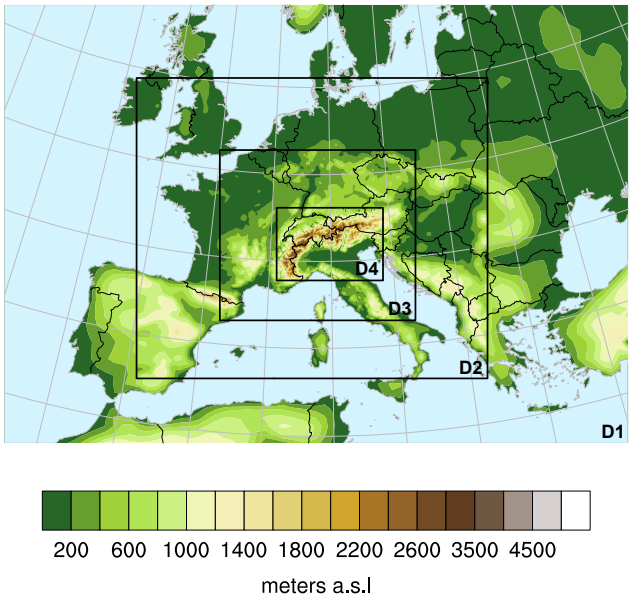


Figure 1. The four nested domains in used in the dynamical downscaling.

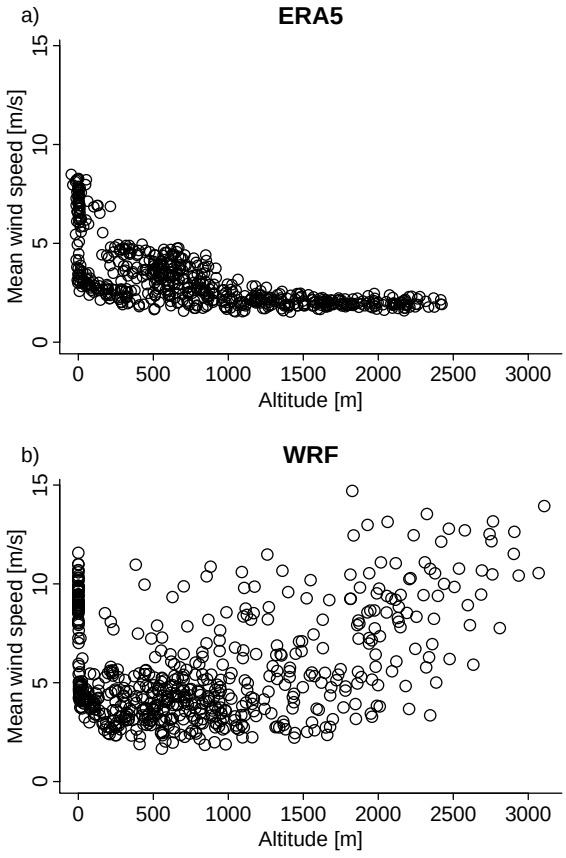


Figure 2. Relationship between mean winter wind speed against altitude for ERA5 (a) and the the WRF model (ERAI-WRF simulation) (b).

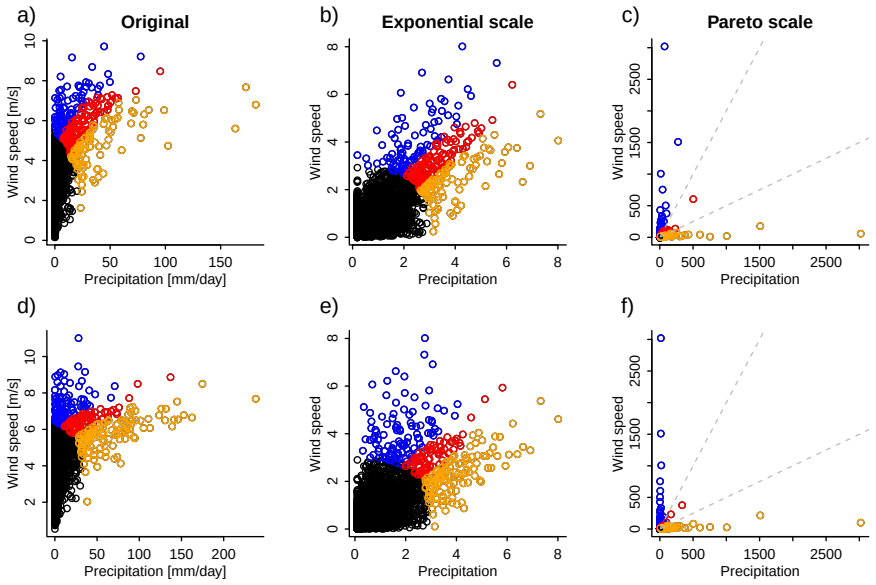


Figure 3. Scatterplots of daily precipitation and wind speed in November-March (1980-1999) for the location with the highest tail dependence χ ($q = 0.95$) in the ERAI-WRF simulations (a-c). CESM-WRF simulations for the same location are shown in (d-f). Shown are the original values (a and d), after transformation into exponential marginals (b and e) and after transformation into Pareto marginals (c and f). The colors highlight the three separating sets W to compute the KL divergence, see Eq. (1), for a high threshold (see main text). In c) and f), the three sets are separated by dashed lines.

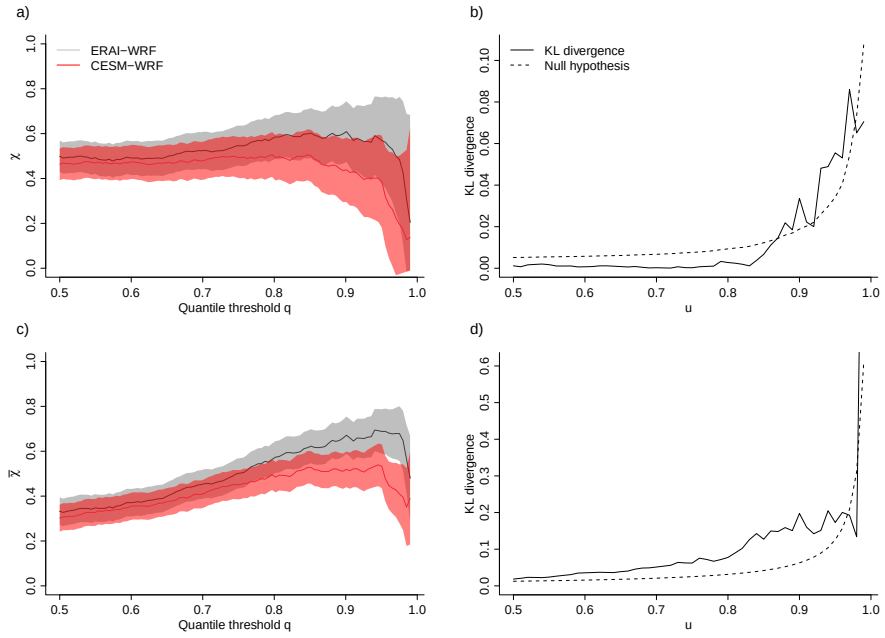


Figure 4. Illustration of the distance metrics between bivariate tails for the location with highest estimated tail dependence χ at $q = 0.95$ in ERAI-WRF. Left: Tail dependence parameters χ (a) and $\bar{\chi}$ (c) for daily precipitation sums and daily maximum wind speed for different quantile-based thresholds q. Shading highlights the 95% confidence intervals. Grey: ERAI-WRF. Red: CESM-WRF. Right: Two different Kullback–Leibler (KL) divergences (eq. (1) with $W = 5$) for the tails of the bivariate precipitation-wind speed distribution between ERAI-WRF and CESM-WRF (solid lines). Dashed lines highlight the 95% confidence interval of the null hypothesis assuming an equal dependence structure. b) KL divergence based on the minimum (i.e., $\min(X_1, X_2) > u$). d) KL divergence based on the sum (i.e., $X_1 + X_2 > u$).

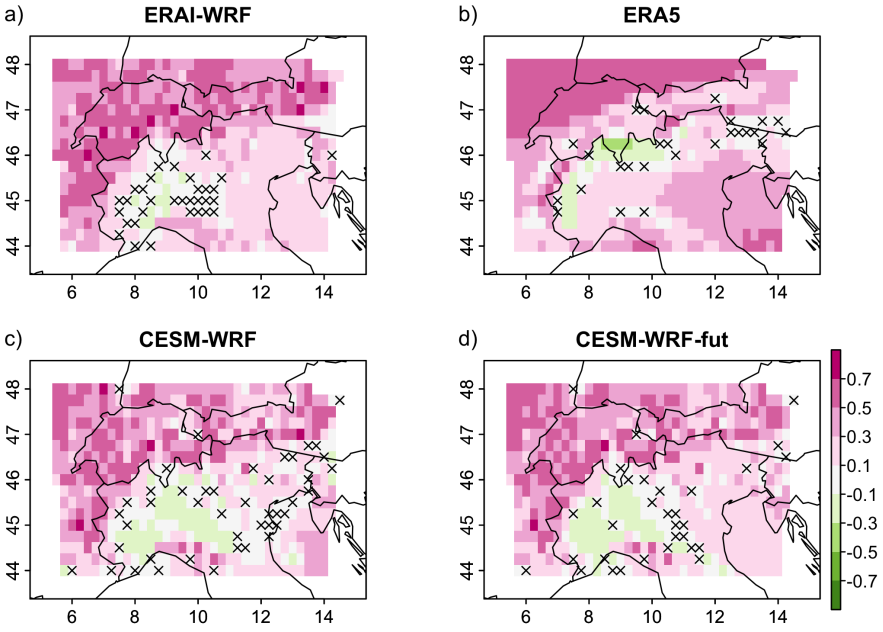


Figure 5. Spearman's rank correlation between daily precipitation sums and maximum wind speed in the extended winter (November-March). a) ERAI-WRF, b) ERA5, c) CESM-WRF, d) CESM-WRF-fut. Non-significant correlations ($\alpha = 0.05$) are marked with a cross.

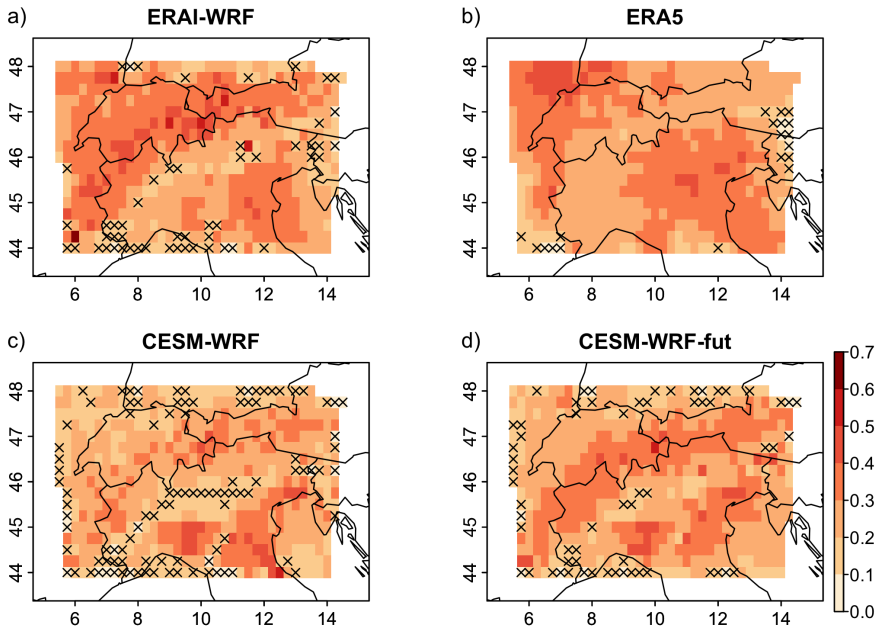


Figure 6. Tail dependence (χ with $q = 0.95$) between daily precipitation sums and maximum wind speed in the extended winter (November-March). Tail dependence was computed considering block maxima over a maximum range 5 days temporally and 1.75 degrees spatially. a) ERAI-WRF, b) ERA5, c) CESM-WRF, d) CESM-WRF-fut. Non-significant values based on bootstraps with the same maximum block size ($\alpha = 0.05$) are marked with a cross.

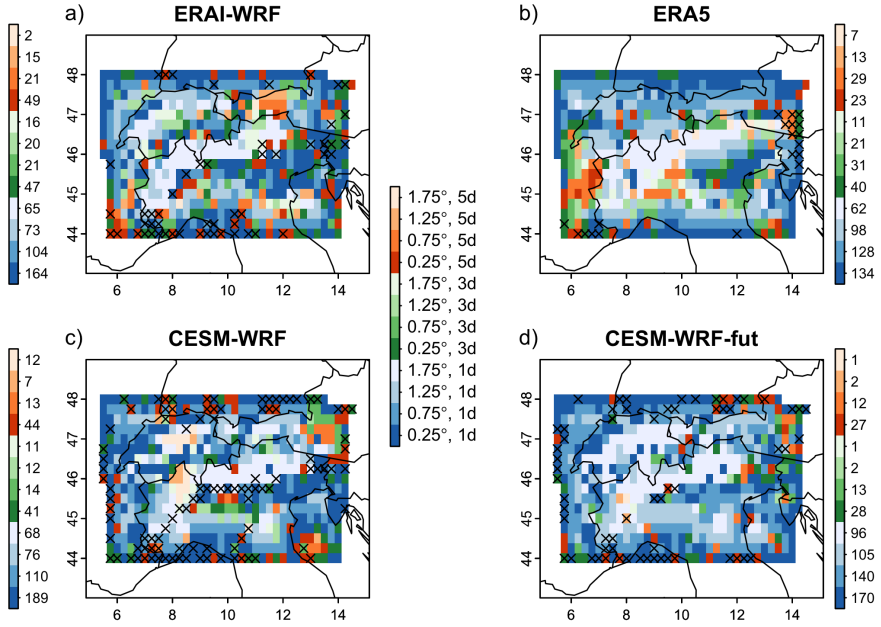


Figure 7. Blocks for which the maximum tail dependence (χ with $q = 0.95$) between daily precipitation sums and maximum wind speed in the extended winter (November-March) is attained (Figure 6). Block sizes range from 0.25°, 1 day to 1.75°, 5 days. Blue, green and orange refer to time lags of 1, 3 and 5 days respectively. Darker shading illustrates higher spatial proximity. The color bars next to the maps show the number of grid points of that color in the corresponding map. a) ERAI-WRF, b) ERA5, c) CESM-WRF, d) CESM-WRF-fut. Grid points with non-significant tail dependence are marked with a cross (see Figure 6).

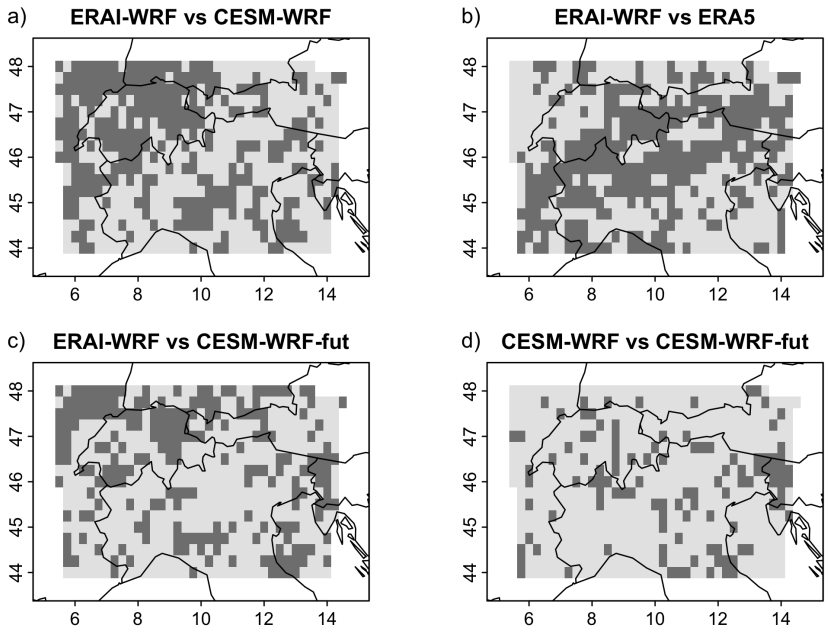


Figure 8. Locations for which the dependence between the tails of daily precipitation sums and wind speed maxima is significantly different based on the KL divergence, Eq. (1) with $u = 0.9$ and ~~$K = 3$~~ $W = 5$ (dark grey, with $\alpha = 0.05$). Dependence is assessed for the blocks that attain maximum tail dependence χ (at $q = 0.95$) (see Figure 6). Shown are comparisons between a) ERAI-WRF and CESM-WRF, b) ERAI-WRF and ERA5, c) CESM-WRF and CESM-WRF-fut, and d) ERAI-WRF and CESM-WRF-fut.

Table 1. Sensitivity analysis of KL divergence (eq. (1)). Reported is the fraction of grid points with significantly different ($\alpha = 0.05$) precipitation-wind speed dependence structure between two datasets for different thresholds u (with ~~$K = 3$~~$W = 5$). The case $u = 0.90$ is shown in Figure 8.

	$u = 0.80$	$u = 0.85$	$u = 0.90$	$u = 0.95$
ERAI-WRF vs CESM-WRF	~~0.40~~ 0.61	~~0.43~~ 0.54	~~0.40~~ 0.46	0.32
ERAI-WRF vs ERA5	0.59	0.53	~~0.47~~ 0.49	~~0.42~~ ~~0.31~~ 0.40
ERAI-WRF vs CESM-WRF-fut	~~0.34~~ 0.53	~~0.31~~ 0.45	~~0.28~~ 0.36	~~0.16~~ 0.27
CESM-WRF vs CESM-WRF-fut	~~0.22~~ 0.30	~~0.19~~ 0.23	~~0.15~~ 0.18	~~0.10~~ 0.19

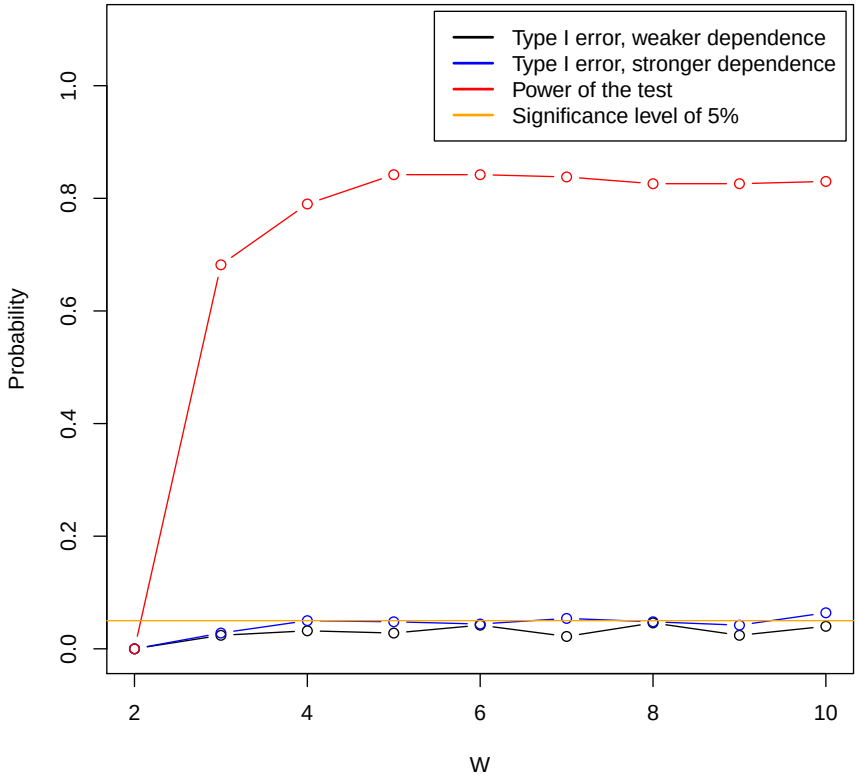


Figure A1. Simulation study using empirical margins.

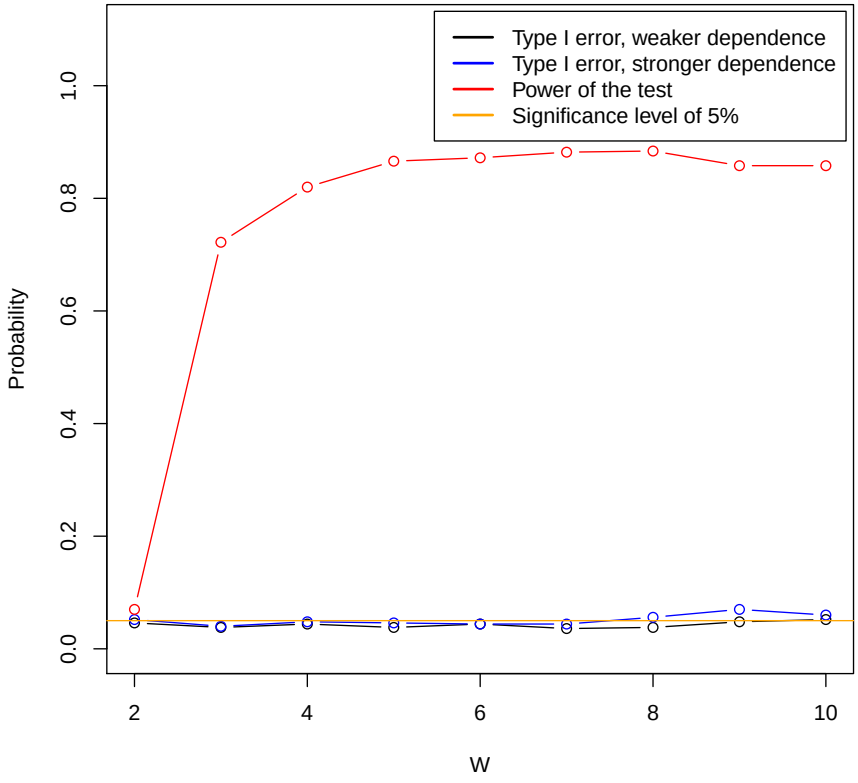


Figure A2. Simulation study using true margins.