Interactive Discussion: Author Response to Referee #1

Present and future trends of extreme short-term rainfall events in Germany, by downscaling convective environments of ERA5 and a CMIP6 ensemble

Gerd Bürger and Maik Heistermann
EGUSphere, doi:10.5194/egusphere-2025-3584

RC: ***Reviewer Comment***, AR: *Author Response*, ☐ Manuscript text

Dear Referee,

thank you for taking the time and effort to review this manuscript. You have managed to dive quite deeply into the modeling setup and have, accordingly, uncovered a number of deficits.

We agree to most of your comments, and will address them in a revised manuscript accordingly.

Please find a point-by-point response to your comments below. These should be considered as preliminary (part of the interactive discussion) since the actual implementation of changes depends on another referee report and the editorial decision.

Thanks again for your efforts!

Kind regards,
Gerd Bürger and Maik Heistermann

1. Comments and responses

1.1. Major comments

RC: ***The manuscript is refreshingly short as it defers many details to elsewhere. Which makes it difficult to review (and reproduce) as information needs to be found elsewhere.***

AR: It was our goal to focus on the projections and their uncertainty. We had hoped that referring the reader to our previous paper Bürger and Heistermann 2023 (BH) would sufficiently cover the explanation of methods, but the reviewer convincingly argues otherwise. Nevertheless, clarifying *all* the raised issues would inappropriately inflate the main text, so we shall divert the additional explanation to the supplementary information (SI).

RC: ***Furthermore, several steps are carried out in a non-standard or non clearly explained way which allows to question the results, e.g. a linear gaussian model for quantiles instead of quantile regression (figure 1), a plethora of neural networks designed for image classification (without giving motivation why they are considered adequate here, table 3), particular implementation of logistic regression (see BH), modelling counts with linear gaussian models instead of Poisson- or similar regression, being not clear on how model climatologies have been obtained, using RPSS in a setting it has not been designed for (nominal***

categories).

AR: Figure 1 was meant as an eye catching and most simple introduction of what the study is all about: that the annual CatRaRE maxima are robustly increasing. Quantile regression is perhaps more compact technically, but poses greater hurdles for the average reader. And visually, moving away from the annual aggregation to the full CatRaRE set would have blurred the picture considerably. For completeness, nevertheless, we plan to provide quantile regression estimates in the revised SI.

The description of the shallow methods was indeed too sketchy and partly inaccurate. This will be addressed in the SI. There we also provide a specification of the reference climate that was missing so far, cf. below.

The reviewer correctly criticizes the description of RPSS. We will fill the gap in the revision, noting already that nominal categories can of course be ordered, e.g. lexicographically as we did, and apply the RPSS in a useful way. See the details below.

The use of deep learning image classifiers had been motivated in BH and shall be so included in the SI.

1.2. Specific comments

RC: ***I can see their setting being similar to statistical downscaling. However, as the target is a categorical variable (1,R,0) which is neither available on the large scale, nor at a small scale and thus could also not be obtained directly from RCMs, I'd like to ask the authors if there is a better name / category for their concept. Something like "impact assessment" or "deriving convective precipitation related indices form GCMs"?***

AR: Here we disagree. The occurrence of a CatRaRE-type event in a specific region, as our target variable, could well be captured in a sufficiently resolved RCM (or, more specifically, in a convection-permitting model). Our approach is equivalent to the weather typing method, as a well-known example of the broader scheme of climate downscaing (e.g. Boé et al., 2006; Conway and Jones, 1998; Gutiérrez et al., 2013).

RC: ***2.1 Methods used and their description: In BH I found with respect to the description of "shallow" models: "As competitive benchmarks to DL models, we employ four shallow statistical models: lasso logistic regression (LASSO), random forests (TREE), a simple neural net with two hidden layers (NNET) and logistic regression based on non-linear least squares (NLS). All of these are applied with and without empirical orthogonal function (EOF) truncation, using North's 'rule of thumb' to find 33, 27 and 21 principal component predictors for cape, cp and tcw, respectively, as estimated from the calibration period; more details are listed in Table 1 and in the source code mentioned at the end." Unfortunately, from this information, I cannot evaluate how the authors actually used these models. With the current pressure on the review system, it is difficult for a reviewer to go to the code and understand what has been done, particularly when the code is not written in the reviewers preferred language.***

AR: An extra section (S2) of the SI now provides a sufficiently detailed explanation.

RC: ***As an example, consider the logistic regression and LASSO logistic regression: In BH, the authors report only the LASSO cost function with the penalty term. Even if we assume that the reader knows the basic idea of logistic regression (probabilities as target and logit-transformation), the following remains open: a) how do the authors deal with the 6 categories? The default logistic regression gives probabilities for two categories (event / no event), how do the authors get probabilities for 6 categories? Do they use multinomial logistic regression? Do they account for the fact that the 6 categories can not be ordered, i.e. they are nominal not ordinal? Or do they use binominal logistic regression for each of the six categories? The latter would not be self-consistent as probabilities for the 6 categories do not necessarily add up to 1.***

AR: Yes indeed, more clarification is necessary. As explained now in the SI, we indeed use multinomial logistic regression, followed by the softmax function to obtain probabilities. – The 6 nominal categories were ordered, actually without much ado, in a lexicographic way (”0”, ”1”; ”NE”; ”NW”; ”SE”; ”SW”). The ordering affects only the calculation of the ranked probabilities, the effect being uniform, however, across methods and simulations. The RPSS, as a relative measure, should hardly be influenced by it.

RC: ***b) why does BH associate a ”non-linear least squares” approach to logistic regression? To my knowledge, logistic regression is a special case of generalised linear models (GLMs) and can be solved in a likelihood framework with iteratively reweighted least-squares (IRLS), see e.g. Dobson and Barnett, ”An Introduction to Generalized Linear Models” (2008). Is the approach given here equivalent?***

AR: It’s not equivalent, as we use a somewhat non-standard approach: for each single class a binary nonlinear least squares optimization via Levenberg-Marquardt is conducted; the results are subsequently combined via the softmax function to obtain the full classification. The approach should be seen as merely heuristic. This is now reflected in the text.

RC: ***c) how does the predictor look like? Do cape, cin, cp for every grid point enter as terms in the predictor? Or do the authors use an EOF as they mentioned in BH? How is this EOF carried out? Are cape fields treated separately from cin and cp as in BH? Or do all three fields enter the EOF simultaneously here? If analogously to BH, cin and cape are seperately. In the current manuscript, cp is used, in BH tcw. Should the reader now assume that cp is treated analogously to tcw? Is the truncation for cp the same as for tcw? Can terms in the predictor of the logistic regression interact? If not, the comparison of logistic regression to random forest would not be fair as for random forest factors interact there by default. To understand how logistic regression has been used here, the reader needs the model equation, e.g. logit(p) cape + cin + cape : cin + ...***

AR: As in BH, EOF reduction, based on North’s rule of thumb, is performed on any of the atmospheric input fields prior to the application of any shallow method; the number of retained EOFs is automatically obtained from the respective eigenvalue spectrum. We clarified the manuscript accordingly.

RC: ***Same for random forest: the reader needs to know what features can be used to define the split rules, are these cape and cin on the grid points or PCs from an EOF? How has that EOF been carried out? How many features can be used at each split? How is the data sampled for each tree? BH gives some information on that but does this transfer directly to the study at hand as we have a different set of predictors and predictant? I assume the reader expects an abstract (i.e. code independent) description of the details of the methods. Imagine that the reader is not familiar with matlab or octave but uses python or R instead. Thus a code independent description is needed. Saying that does not mean that it is not useful to publish the code!***

AR: As in BH we use a default setting of M5PrimeLab after EOF filtering; this implies to use a third of the input variables at random with replacement. From its author G. Jekabsons we have learned that, by default, M5PrimeLab uses only linear predictor terms. Manuscript text has been augmented accordingly.

RC: ***I refrain from discussion the other methods in this way. Also because the neural network based methods can not be described in the way a logistic regression model can be decribed. I expect, however, that the reader needs an argument why a given method has been chosen for the canon of methods, e.g. giving typical usecases of the networks and a short explanation why this should be useful in the case at hand. I can imagine that image classification is similar to the case we have here with the three maps (cape, cin, cp). Also color images have a grid structure with three values (RGB) at each grid points. So I see the similarity. Doing some research on LeNet-5, I find that this is used for very special images, namely of size 28x28,***

greyscale. Why is that considered here?

AR: This point was covered in greater detail in BH and is now, in an extra paragraph, also covered in the revised manuscript.

RC: ***The ANOVA in Sec. 3.3 includes as factors GCM, region, scenario, classifier and severity. While I see GCM and classifier (and maybe region) as interesting factors to understand the uncertainty in the trend values, I do not understand why the authors use severity and scenario as factors. I'd expect an ANOVA using GCM and classifier (and maybe region) separately for various severity classes and scenarios to obtain an idea about the uncertainty for a given severity. Including severity as a factor seems like quantifying the uncertainty for trends of two different variables, such as for temperature and precipitation. Please correct me if I am wrong here.***

AR: You certainly have a point here. But along that same line one could argue that the NE trend is a different variable than the SW trend. We prefer to think of trend variability (variance) instead of uncertainty, starting from complete ignorance. This includes, for example, the view that emission scenarios may not have any significant influence on the trends at all, very much in the spirit of Lehner et al. (2020).

1.3. Technical corrections

RC: ***Title: "... downscaling convective environments ..." are really the convective environments (characterized by cape, cin, cp) transformed to a smaller scale? From my point of view, the study rather derives an impact related index directly from a GCM without dynamical or statistical downscaling.***

AR: True. But as mentioned above, it is a downscaling of extreme rainfall characteristics.

RC: ***l. 14: 70B EUR → better 70 billion (most natural in running text) or €70bn or €70 bn ("bn" is preferred over a capital "B")***

AR: Corrected.

RC: ***l. 17: "evidence of this ...": "this" is related to what?***

AR: Corrected.

RC: ***l. 49 (and many other occasions): reading this out loud "99%-percentile" sounds strange as there is twice "percent". Suggestion: 0.99-quantile, 99%-quantile or 99. percentile. For higher quantiles use 0.999-quantile or 99.9%-quantile and 0.9999-quantile or 99.99%-quantile, respectively. Percentiles are associated with quantiles specified by integer percentages. The use of non-integer percentiles can be frequently seen but does not fit to the concept.***

AR: Corrected.

RC: ***l. 49: Are the trends obtained using quantile regression? If not I would be scepticle about the p-values. If a gaussian linear model is fit to quantiles, the assumptions might not fit and inference is not robust. Furthermore, the quantiles are themselves estimates with an uncertainty which should be accounted for. A gaussian linear approach still might give sensible estimates in some cases but quantile regression is definitively more appropriate here. As this is technical very simple, I suggest the authors use that to increase confidence in their study.***

AR: As argued above, we prefer to deal with annual statistics here because they are more accessible to the average reader than quantile regression, but include a passage on quantile regression in the new SI. See also our next

response.

RC: ***l. 53: The authors argue that their trends are quite robust. What is meant here by robust? Removing some points lead to similar trends?***

AR: We have added an extra paragraph as follows:

> We note that throughout the following, trends are estimated using linear (ordinary least squares) regression, and that in all cases corresponding residuals are consistent with this model (normality), so that trend lines and significance estimates are reliable. In the case above this is quite remarkable because, after all, we are looking at a measure of annual extremes of precipitation; it is a likely consequence of the integrating effect of that measure, the extremeness index $E_{T,A}$.

RC: ***l. 54-55: "... corresponding to the overall, 90% and 99% ..." the missing comma after overall seems very important***

AR: There is no comma... anyway, we have rewritten the sentence as follows in order to make it more unambiguous:

> To further study these very extreme cases, we select events that have $E_{T,A} \geq 8.7$ and $E_{T,A} \geq 20.7$, corresponding to the 90 and 99 percentiles and representing the 2040 and 208 strongest cases, respectively. The severity classes, denoted as P_{00} (all events), P_{90}, and P_{99}, respectively, should be sufficiently large to undergo the statistical modeling as described below.

RC: ***l. 72: "As cape and cin are not provided by the CMIP6 models they had to be calculated from these profiles, following Bolton (1980)" → give more details on that in the supplementary material or an appendix.***

AR: Added in the SI (S3).

RC: ***l. 79: why is the 10x10 grid used? This I did not understand. Maybe it has been explaind in BH? But the reader needs some information here.***

AR: We have now revised the manuscript as follows:

> Spatially, we allowed only models with an original resolution not coarser than 1.5×1.5 degrees, and interpolated or aggregated their fields to a unique 10×10 grid between the corners [5.75E 47.25N] and [15.25E 55.25N].

RC: ***l. 82: "reference climate" this is not sufficient. Is the climatology obtained per day of the year? Per month? Or for every of the four time steps a day has (my favourite)? Or is it a mean over all time steps in the data set? Does every grid point has its own climatology (my favourite)? The text suggests that there is an average over all grid points. An equation would be unambiguous here.***

AR: As the text suggests. For clarification, the sentence now reads:

> For any GCM and field variable ξ, the reference climate was calculated as mean and standard deviation (μ, σ) across all grid points and historical realizations of the same GCM from the reference period 1981–2010; from that, standardized anomalies (ξ - μ)/σ are formed to enter the classification.

We understand and actually share the reviewer's preference, but given the relatively short reference period and corresponding sampling uncertainty we had opted for the simplest option here.

RC: ***l. 91: What is meant by "probabilistically classified"? The classification does also need some more explanation. There are 6 nominal classes, i.e. they cannot be ordered. This has consequences on a) the logistic regression approach (multinomal regression) and b) the verification: RPS is not adequate as the classes (NE, NW, SE, SW) have no natural ordering but that is what the ranked probability score builds upon. → comment later.***

AR: We removed *"probabilistically"*. Otherwise see our responses above and below.

RC: ***l. 102: I find the term "shallow" strange for a random forest and logistic regression as it puts these in a the framework of neural networks without any need.***

AR: You certainly have a point. But first we wanted to remain close to the nomenclature of BH, and second we emphasized the distinction between 'conventional' and the presently quite popular approaches around AI, which usually means 'deep' as we use it here.

RC: ***table 3: "Nonlinear least squares" as solver for the cost function related to logistic regression sounds strange, see my comments in 2.1. Why are there no number of parameters associated with the "shallow" models? If the authors put random forests and logistic regression in the frame of neural networks, why not giving the number of layers? LASSO, TREE and NLS have 1, NNET hast 2, right? Why has "logreg" 32x32 nodes? I do not find an explanation in BH, neither in the associated supplement.***

AR: Regarding NLS, see our comment above. Furthermore, we revised the column headers: the number of layers and parameters now only applies to the DL methods, in order to estimate the overfitting potential (with #Params being several orders of magnitude larger for DL).

For the 32x32 resolution, the caption reads now:

> Table 3. The conventional ("shallow") and deep learning classifiers. The DL resolution is taken from corresponding images of typical applications, e. g. 32x32 from the CIFAR-10 dataset.

RC: ***l. 106: Exceedance counting: It is not clear what this chapter is good for.***

AR: The section is meant as a bridge between the actual CatRaRE events shown with corresponding E_{TA} values in Fig. 1 and the class probabilities that can be derived with our approach. For clarification, we added the sentence:

> But such a scheme is incapable of deriving actual events with concrete values of $E_{T,A}$, as displayed in Fig. 1. To relate class probabilities to Fig. 1, we show in Fig. 3 for the severity classes $P_{90}, P_{99}, P_{99.9}$ (i.e. the all-time 90-, 99-, and 99.9-percentiles) the corresponding annual exceedance counts. These quantities, in particular their centennial trends, are the target quantities for which this study provides future estimates. In Fig. 3 they are shown relative to their overall mean per severity class, to allow a comparison across these classes.

RC: ***l. 107: "zero-order projection", as you extend the trend, I would call it a first-order projection.***

AR: We removed the term entirely.

RC: ***figure 3 / l. 112: are these quantiles (percentiles) those resulting from the fit in figure 1? Or are these the overall temporal quantiles? What are the annual exceedance counts here? In l. 112 they are called "annual relative exceedance counts normalized by the overall exceedance count". This is not fully clear to me. I suggest an equation. This figure needs some more explanation. As these are counts, Poisson regression would be an appropriate trend model. Is this used here? Otherwise information on significance might be unreliable.***

AR: See our previous response. As an example, take P_{90} to be the 90%-quantile of E_{TA} for the full time series, and let for each year y $e_y = \#(E_{TA} \geq P_{90})$ be the number of events for that year exceeding P_{90}. Each dot in Fig. 3 represents a quantity $e_y/\overline{e_y}$, where the mean is taken over all the years. The text is now:

> To relate class probabilities to Fig. 1, we show in Fig. 3 for the severity classes $P_{90}, P_{99}, P_{99.9}$ (i.e. the all-time 90-, 99-, and 99.9-percentiles) the corresponding annual exceedance counts, divided by their overall mean.

As the data are scaled that way, standard (OLS) regression should be sufficient for this purpose.

RC: ***l. 122: "cross-entropy" needs a reference, best for a similar case.***

AR: We now have:

> For the DL training, cross-entropy loss (equivalent to logistic loss, see, e.g., Goodfellow et al., 2016) is used as a loss function.

RC: ***l. 123: "Ranked Probability Skill Score (RPSS)" is the skill score based on the Ranked probability score, a score for probabilistic information for ordinal categories. Here we cannot order the categories and thus I doubt that the score is appropriate. Furthermore, the authors need to specify the reference prediction as needed for every skill score. Is it the climatology? I would be interested in seeing the reference for "RPSS is the skill score corresponding to the relative deviance of the predicted probabilities from the actual outcome".***

AR: We assume that when you say the categories cannot be ordered you mean they cannot be ordered in a physically sensible way, which is probably right in our case. We recall that the RPSS was introduced in the context of *propriety*: If a physically meaningful ordering exists, the original multinomial probability score, PS (Brier, 1950; Murphy, 1966), turns out to be *improper*, which motivated Epstein (1969) to introduce the *ranked* probability score (RPS) as a *proper* extension of the PS. But this does not mean, conversely, that for a non-physical ordering, such as lexicographic (see above), the RPS is not useful. It is still a very useful score for multinomial probabilistic predictions, similar to the original PS, even more so as the RPSS is a relative score, as mentioned above. With respect to reference prediction and source, the text is now modified as follows:

> RPSS is the skill score corresponding to the deviance of the predicted probabilities from the actual outcome, relative to predicting climatology, cf. Wilks (2006, p. 302)

RC: ***figure 4: RPS, see above. Why are only 12 methods shown? Table 3 gives 13 methods, however, it is not clear what is the difference between "Logistic regression (NLS)" and "Logreg (2023)". The latter is not***

showing up in figure 4. Why is there a range for RPSS? Is there a variation of model hyperparameters? Please explain this in more detail.

AR: The caption reads now:

> Ranked probability skill score for the skillful classifiers, based on all CatRaRE events (a, P00) and on the upper extremes P90 and P99 (b, c). For visibility, classifiers with no skill (Logreg, LASSO) are skipped.

With regard to the skill range, please note the new sentence which we have inserted near l. 135 of the preprint:

> We recall from BH that for DL there is a range in skill due to the stochastic nature of the optimization;

RC: ***l. 141: Please give the interpretation of the "averages of the resulting 6-hourly probabilities", e.g. expected number of events per 6h interval. Maybe it would be better to work with the sum over a summer period to give expected numbers for the summer?***

AR: Each simulated value represents the probability of an event happening during the corresponding 6 hour interval in the respective class, cf. §2.3. From that we form, for any class and year, the corresponding average over all such probabilities.

RC: ***l. 152: "... the 20 entries of the table ..." which table?***

AR: These are the 20 realizations (r1, r2, ...) of Table 1. We clarified this in the text.

RC: ***l. 160: if trends for probabilities are not obtained with logistic regression, I would be sceptical with the significance.***

AR: The errors about the trend line do not seem to contradict the assumptions of linear OLS regression (independence, etc.). And normality of residuals cannot be rejected based on the Kolmogorov-Smirnov test, as mentioned above.

RC: ***l. 166: "... analogous to 7are Figs...." -> "... analogous to Fig. 7 are Figs...."***

AR: This has been corrected.

RC: ***figure 6: here are trends obtained for probabilities → use logistic regression. "Annual mean probabilities FOR class 1 ...". Explain the right column of the plot. How is OBS obtained?***

AR: For trends, see our remark to l. 160. With respect to OBS, we have added that it is obtained as annual relative class 1 frequency of CatRaRE. The right column is exactly like the left but confined to the common period, and the time axis is replaced by the OBS values.

RC: ***figure 8: Are these cape, cin and cp values averaged over the whole summers? Is this meaningful? It is short term strong cape which favours convection and cape is reduced through convective precipitation, so the mean does not need to be high but there can still be convective precipitation.***

AR: Your argument certainly applies to a process-based view. But here we inspect long-term behavior, and for that it is plausible to assume that the effect of cape depletion from convective precipitation remains unchanged in the long run, so the summerly average should be ok. For clarification we have augmented the text now as follows:

To study their long-term behavior, we again form annual averages and corresponding trends, assuming implicitly that diurnal effects between them, such as cape depletion from convective precipitation, remain unaffected by climate change.

RC: ***l. 185: "The strong skewness for SSP126 may be related to the cape vs. cin antagonism" → Could that not be checked by looking at individual cases?***

AR: Such an approach might be a helpful entry point; still, a more comprehensive approach would be required to obtain a conclusive answer - an approach that we consider beyond the scope of the present study.

RC: ***l. 204: "Fig. 12 shows the ANOVA based on significant trends only" What is the argument for using significant trends only? The insignificant ones are either small and yield values close to 0, which would be ok for including them in the analysis, right? OR they are based on a small number of events (high severity class). This effect should be taken care of anyway by using only one fixed severity classes for an ANOVA.***

AR: We argue that the large number of insignificant trends outweighs the significant ones. It can thus produce an effect on the ANOVA results, and it indeed does. We must admit that we discussed this issue controversially without a conclusive result. We deemed it the best solution to show both ANOVA results - with and without insignificant trends - as both allow for specific insights.

RC: ***l. 214: "... aspects of the modeling chain ..." Yes, I agree. There are however more aspects to modelling chain presented here, which has not been considered, e.g. extremeness index, estimating quantiles or non-optimal regression approaches.***

AR: The relevance of these factor is undisputed. However, we consider uncertainty in them as minor when compared to, for example, GCM or scenario uncertainty, and we preferred not to inflate the presentation of the ANOVA further. However, we mention now in the revised version of the manuscript these additional aspects.

RC: ***l. 217: "... found growing levels of both antagonistic drivers cape and cin, ...". I do not see why the mean of these quantities over the whole summer should give arguments for that. I can imagine constant means over several summers but varying convective precipitation in each summer. As said earlier, we can have high cape which is eaten up by conv precipitation within a day.***

AR: Please see our response above. Note also that e.g. Ukkonen and Mäkelä (2019, Fig. 6) demonstrate the lagged effect of cape on rainfall based on spatial and long-term diurnal averages. We agree that averaging CAPE over a full summer means to discard information about the actual temporal distribution with this period. In our view, however, it is obvious that a higher mean value implies a different distribution, including more days with very high CAPE values. But while we think that looking at the mean summerly CAPE over very long time periods should be indicative of the gradual changes in convective atmospheres, we agree that further analysis appears warranted and consider to elaborate on this in the revised manuscript.

Note, however, that using e.g. higher quantiles instead of the mean (not shown) does not change the picture much.

RC: ***l. 227: "For downscaling this may be good news, as it means that very different approaches still arrive at similar results." Is this not in contradiction with your RPSS figure? From my understanding, this means, that the variability due to different GCMs is just much larger than the differences between classifiers. Do you really want to have this classification method related sentence as last sentence of your conclusion? Does that fit to your title? Should not a sentence about the trends at the end?***

AR: Indeed, we agree with the referee's claim that "the variability due to different GCMs is just much larger than the differences between classifiers." However, we do not see the contradiction to the RPSS-related figures, since these exactly show that the actual differences in the performance of the various classifiers is not high. Regarding the last sentence we have basically reversed the ordering, so that the final paragraph now reads:

> With respect to uncertainty in the centennial trends, the choice of classifier (downscaling) is less relevant, and so is the region considered. For downscaling this may be good news, as it means that very different approaches still arrive at similar results. For the regions, heavy precipitation, in response to global warming, appears to be increasing uniformly across all of them, and that, if news at all, is bad news anyway. The large uncertainty around these results mainly stems from the choice of climate model, followed by emission strength. This is reversed if insignificant trends are ignored, resulting in a largely reduced ensemble but with emissions now becoming the main factor for determining the trend.

References

Boé, J., Terray, L., Habets, F., and Martin, E.: A simple statistical-dynamical downscaling scheme based on weather types and conditional resampling, Journal of Geophysical Research: Atmospheres, 111, https://doi.org/10.1029/2005JD00688 2006.

Brier, G. W.: VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY, Monthly Weather Review, 78, 1–3, https://doi.org/10.1175/1520-0493(1950)078

Conway, D. and Jones, P. D.: The use of weather types and air flow indices for GCM downscaling, Journal of Hydrology, 212–213, 348–361, https://doi.org/10.1016/S0022-1694(98)00216-9, 1998.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, Academic Press, 2006.

Epstein, E. S.: A Scoring System for Probability Forecasts of Ranked Categories, Journal of Applied Meteorology and Climatology, 8, 985–987, https://doi.org/10.1175/1520-0450(1969)008

Goodfellow, I., Bengio, Y., and Courville, A.: Deep Learning, MIT Press, 2016.

Gutiérrez, J. M., San-Martín, D., Brands, S., Manzanas, R., and Herrera, S.: Reassessing Statistical Downscaling Techniques for Their Robust Application under Climate Change Conditions, Journal of Climate, 26, 171–188, https://doi.org/10.1175/JCLI-D-11-00687.1, 2013.

Lehner, F., Deser, C., Maher, N., Marotzke, J., Fischer, E. M., Brunner, L., Knutti, R., and Hawkins, E.: Partitioning climate projection uncertainty with multiple large ensembles and CMIP5/6, Earth System Dynamics, 11, 491–508, https://doi.org/10.5194/esd-11-491-2020, 2020.

Murphy, A. H.: A Note on the Utility of Probabilistic Predictions and the Probability Score in the Cost-Loss Ratio Decision Situation, Journal of Applied Meteorology and Climatology, 5, 534–537, https://doi.org/10.1175/1520-0450(1966)005

Ukkonen, P., and A. Mäkelä: Evaluation of Machine Learning Classifiers for Predicting Deep Convection,Journal of Advances in Modeling Earth Systems 11(6), 1784–1802, 2019.