Nonlin. Processes Geophys. Discuss.,
https://doi.org/10.5194/npg-2019-62-AC2, 2020


Nonlinear Processes in Geophysics
Discussions

# *Interactive comment on* "Simulation-based comparison of multivariate ensemble post-processing methods" *by* Sebastian Lerch et al.

**Sebastian Lerch et al.**

sebastian.lerch@kit.edu

Received and published: 2 May 2020

*We thank the three reviewers for their positive assessment and their thoughtful comments which, we believe, will strengthen the manuscript. Below we addressed each comment in turn. Our replies are in italics. We also created a track-changes PDF for your convenience.*

*Kind regards,*
*Sebastian Lerch, Sándor Baran, Annette Möller, Jürgen Groß, Roman Schefzik, Stephan Hemri and Maximiliane Graeter*



In this paper, a comprehensive simulation study is implemented, comparing multiple multivariate statistical post-processing methods which all combine standard univariate post-processing with one of four techniques to reintroduce spatial, temporal or inter-variable dependencies. It is well-written, an important contribution given the variety of techniques available and potentially very useful to identify optimal operational post-processing strategies for varying types of data. Just a few clarifications and changes are needed.

General comments:

I would have liked to have more focus (or at least comments) on the effect of ensemble size and dimension. The here chosen ensemble of 50 members is at the upper limit of what is now operationally produced, with the majority far below this number. Of course it is important for the study to have a sufficient number of data points so as to produce significant results, but it would also be interesting to look at settings with a smaller number of ensemble members. Also,the number of dimensions ranges from 4 to 5, which would correspond to looking at consistency between a few weather variables, but would usually be too low for a setting where preserving spatial or temporal features are important. I wonder if the findings would be different for smaller ensembles or higher dimensions.

*Thank you for this suggestion. Following a similar comment by Reviewer 1, we performed additional simulations to assess the effects of ensemble size. Please see the response to Reviewer 1 for a discussion of the role of the number of ensemble members.*

*To study the effects of the number of dimensions, we performed additional simulations for Setting 1 with dimensions between 2 and 50. Overall, the relative results are often not effected too much by changes in the number of dimensions, in particular in terms of the energy score. Somewhat more substantial differences can be observed*

*in terms of the variogram score. In a nutshell, GCA performs worse for higher dimensions, whereas ECC-S improves for larger values of $d$. The relative differences to SSh (in favor of SSh) become more substantial with increasing $d$, whereas the changes in relative performance of dECC show a strong dependency on the combined misspecification of variance and correlation of the raw ensemble.*

*We have added a paragraph to Section 4.2.1. Additional figures with results for $d = 2, 3, 10, 20, 30$ and 50 have been added to the Supplemental Material (Section 1.2 there). See the response to the corresponding comment by Reviewer 1 for a discussion of the display as multiple boxplots within one figure. Performing additional simulations to investigate different choices of $d$ for Setting 3 (which is Setting 2 in the revised manuscript) was impossible due to constraints regarding the computational requirements, see also our comment below.*

I find it very interesting that the performance of certain methods is sometimes very different when $\rho > \rho_0$ than in the opposite case. Do you have any explanation for this?

*We believe that despite the (relative) simplicity of the chosen simulation setup, interpreting differences in multivariate performance is challenging due to inter-connected contributions of misspecifications in mean, variance and correlation structure and their effects on the (still) not well understood multivariate proper scoring rules. We are thus unable to provide a comprehensive explanation for the particular role of the correlation parameters $\rho$ and $\rho_0$. Some potential explanations of differences in forecast performance may be given by observing that under certain circumstances, modifying a raw ensemble prediction by artificially weakening correlations, for example by randomly permuting ensemble member's forecast vectors, may unexpectedly improve predictive performance. This was for example observed in Schefzik (2017), and is likely an issue of underdispersed univariate forecast distributions that improve by the changes on the univariate forecast distributions imposed by the modifications of the correlation structure. However, to assess the effects of different sources of misspecifications in*



*further detail, additional multivariate verification techniques such as multivariate rank histograms should accompany the analysis. While such additional studies are beyond the scope of the paper, they represent an interesting starting point for future research.*

Specific comments

1. Line 71 and Line 153: The correlation matrix here is not necessarily the identity matrix, so I don't think it is a standard normal distribution.

*Fixed.*

2. Line 74: I would mention here that $m$ is the number of ensemble members

*Added.*

3. Section 2.2: There is a mixture of $x$ and $X$ used to define samples and ensemble forecasts, but I was confused why this distinction is made within the notation, it seems inconsistent.

*Throughout the description of the methods, we used $x$ to denote univariate quantities in the individual dimensions. $X$ in bold print is used to represent vector-valued quantities, and $X$ in normal print is used to for components thereof. A sentence has been added at the beginning of Section 2.2 to clarify this.*

4. Lines 184-186: For the other settings it is mentioned to which weather variables these settings could apply. It would be nice to add something like this to Setting 1, as well.

*A corresponding sentence has been added to the beginning of Section 3.1.*

5. Line 242: The notation in this setting is different from the others and this is confusing. Here, the forecasts and observations are marked with $x$ and $y$, whereas the other settings use $o/0$ to mark the observations.

*The notation has been changed accordingly, see also our answer to Reviewer 1.*

6. Line 276: Is there a specific reason, why $d$ is 4 in this setting and 5 in the others?

*The limitation to $d = 4$ was mainly due to computational requirements and numerical stability issues caused by the `NORTARA` package for `R` used to generate the multivariate ensemble predictions and observations. A sentence has been added in the paper.*

7. Lines 292-293: Some of the matrices are in bold face, some are not.

*Fixed. Additional minor corrections to the description of Setting 4 (now Setting 3A) are detailed under "Further changes".*

8. Figure 1: I am a bit surprised to see that GCA is performing that much worse as compared to the other post-processing methods (in a univariate sense). There are even cases where the performance is equal or possibly worse than for the raw ensemble. Do you have an idea why that could be?

*In particular when compared to ECC-Q and related methods, we believe that this is mostly an issue of sampling. As discussed in Section 4.1, the univariate quantile forecasts of ECC-Q are (close to) optimal in terms of the CRPS, but the random samples from the predictive distributions obtained in GCA are not (see next comment). We have modified the corresponding sentence to make this more clear. Cases where performance is worse than the raw ensemble are very rare and likely arise when simulation parameters of the ensemble forecasts are close to these of the observations.*



9. Lines 328-330: Can you explain a bit further what you mean by "optimal in the terms of the CRPS"?

*When the CRPS is used for evaluating a forecast distribution represented by a finite simulated sample, the sample is implicitly interpreted as a set of quantiles from the underlying forecast distribution at levels $\frac{i-0.5}{m}, i = 1, \ldots, m$. Viewed differently, this implies that utilizing quantiles (at these levels) when generating a univariate sample from a forecast distribution will result in a lower expected CRPS than drawing samples at random. Even though the levels at which quantiles are obtained are not identical to the optimal quantile levels, the differences will be small for $m = 50$. Details on the mathematical background can be found in the referenced paper by Bröcker (2012). We hope that this becomes more clear with the modifications made to this part of the paper mentioned in the response to the preceding comment.*

10. Lines 334-335: Naturally, scenario D has the smallest improvement compared to the others. Does that also mean that the scenarios are on the same absolute skill level after post-processing?

*We did not further investigate univariate performance since we removed this paragraph following a suggestion of Reviewer 1.*

11. Footnote 4: In my opinion it would be clearer if you refer to ECC-Q as EMOS-Q in this section as well.

*We have modified the corresponding paragraph. Together with the new table comparing key features of multivariate post-processing methods (see comment by Reviewer 1), we hope that this sufficiently clarifies terminology.*

12. Lines 415-430: Can you refer to the figures in the appendix that show these results by number?

*References to the corresponding sections of the Supplemental Material have been added.*

13. Line 446: “the VS might be better able to account..” This is confirming a known result, therefore “might” is a bit unsuitable.

*Changed as suggested.*

Technical corrections

1. Line 327: I would move the sentence beginning "Note that" to footnote 4, as it directly relates to the changes in the marginal distributions mentioned there.

*The corresponding paragraph has been modified following comments from Reviewer 1, and the footnote has been removed.*

2. Line 356: I find this sentence a bit confusing. Should there be a comma before “the less information”?

*We have modified the sentence and added a comma as suggested.*

3. Line 386: Missing comma before “where”

*Fixed.*

4. Lines 415 and 441: I would add “parameter” after “observation location”.

*Changed as suggested.*

---

Interactive comment on Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2019-62, 2020.