Nonlin. Processes Geophys. Discuss.,
https://doi.org/10.5194/npg-2019-62-AC1, 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

Reviewer 1

This paper reports the results of a comprehensive simulation study comparing different methods for modeling spatial, temporal, and inter-variable correlations in statistically postprocessed ensemble forecasts. With a number of new multivariate methods having been developed in recent years, a study like this is of great interest as it allows readers to get an overview of the strengths and limitations of the different approaches.

General comments:

1. With the goal of this paper being a comparison between different methods for multivariate ensemble postprocessing, I feel that a more detailed discussion of the key features (e.g. optimal sampling of the predictive distribution vs. random sampling, assumption of stationarity of the copula structure vs. flow dependent copula structure, etc.) of the different methods should be given (possibly in the form of a table). This could serve as a motivation for the different simulation settings, which try to mimic situations where some of the assumptions are met while others are not.

Thank you for this suggestion. We have added a table at the beginning of Section 2.2 where we now compare several key features of the different multivariate post-processing methods. We further refer to the related findings in Wilks (2015) (next comment) at the beginning of Section 2.2 and in the Discussion.

2. Related to 1., I feel that the role of ensemble size (which has a big impact on the representation of the multivariate distribution) should be discussed a bit more. This seems relevant as for some methods it is easy to generate an ensemble of any size while for others it is not. I'm not suggesting that additional experiments should be performed, but a brief discussion of the findings in Wilks (2015) could be useful in a context

where strengths and limitations of different multivariate postprocessing approaches are compared.

We agree that the role of ensemble size is important and warranted a more extensive discussion. We have performed additional simulations for Settings 1 and 3 (Settings 1 and 2 in the revised paper) with ensemble sizes between 5 and 100. For Setting 1, the results are largely as it can be expected: The relative differences in terms of the ES between ECC-Q and ECC-S, and between ECC-Q and GCA become increasingly negligible with increasing ensemble size; likely due to increasingly smaller intervals from which random quantiles are drawn. Further, SSh shows improved predictive performance for larger numbers of ensemble members for $\rho_0 < \rho$ in case of the ES, and for for $\rho_0 > \rho$ in case of the VS. This can likely be attributed to the corresponding increase in sample sizes when determining the dependence templates, similar to the effects on GCA. The relative performance of dECC is strongly effected by changes in m for large misspecifications in the correlation parameters. A positive effect of larger numbers of members relative to ECC-Q in terms of both scoring rules can be detected for $\rho_0 > \rho$ when $\sigma < 1$, and for $\rho_0 < \rho$ when $\sigma > 1$. In both cases, the corresponding effects are negative, when the misspecification in σ is reversed.

In Setting 2 (old Setting 3), in contrast to Setting 1, GCA seems to benefit most from an increasing number of members, while SSh benefits only slightly in terms of the ES, and stronger in terms of the VS. This is for example illustrated in the attached Figure 1, which corresponds to the additional scenario presented in Figure 5 in the main paper. As in Setting 1, ECC-S becomes more similar in terms of the ES to the reference ECC-Q for increasing number of members. However, this effect, is not (or not that strongly) observable for the VS, where the number of members has nearly no effect on ECC-S.

We have added a paragraph to Sections 4.2.1 and 4.2.2 where the effect of ensemble size is discussed shortly. Figures with additional results have been added to the Supplemental Material in Sections 1.1 and 2.2 there. We are aware that displaying

boxplots for several choices of m within a single figure is not optimal since the underlying random samples within each panel will necessarily differ across different values of m. Nonetheless, we believe that the differences due to random sampling are negligible due to the application of DM tests and the consideration of 100 repetitions of each simulation experiment. Therefore we chose the illustration added to the Supplemental Material in Sections 1.1 and 2.2 because the effects of changes in m are straightforward to compare.

3. Why has so much focus been given to simulation settings that are based on a time-independent model for the simulated forecasts and observations? I would argue that this (time-independence) is not a feature commonly encountered in applications, but with 3 out of 4 settings being time-independent, multivariate methods that assume a stationary copula structure could be perceived as being more versatile than they really are. Agreed that setting 1 is a natural starting point for such a comparison and that set-ting 3 is interesting because of the entirely different nature of the marginal distributions(skewness, possibility of heavy tails, mixed discrete-continuous distributions), but what do we learn from setting 2 that we cannot learn from 1 and 3? The main difference to 1 seems to be the poorer performance of GCA, but an explanation for this is not given,and so the insights gained from this setting are limited.

We agree that the results of Setting 2 were overall very similar to those of Setting 1. Therefore, we have removed Setting 2 from the paper and added a slightly adjusted version to the Supplemental Material. The former Setting 2 is there denoted by Setting S1, and setup and results are now discussed in Section 5 of the Supplemental Material. Accordingly, the paper has been adjusted as follows:

- *Setting 4 is now Setting 3, and Setting 3 is now Setting 2*
- *Sections 3.2 and 4.2.2 have been removed from the paper and added to the Supplemental Material*

- *Figure 1 has been adjusted by removing results for Setting 2*
- *several references to Setting 2 in the text have been removed throughout*
- *A short paragraph referring to the additional setting (now in the Supplemental Material) has been added to the end of Section 3.1.*

Specific comments:

131-132: Please check if that statement is correct. In my recollection the selection of past observations in Clark et al. (2004) was not random, but was based on the valid date of the forecast

Thank you for pointing this out, you are of course correct. The description of SSh has been corrected and information about the random selection of training cases has been added to Section 3.1 in step (S3).

293-297: I find setting 4 the most interesting, but I find the particular definitions of the model parameters unnecessarily complicated. Specifically, I don't get a good intuition of what kind of time-varying correlations this model implies. Couldn't one simply define

$$\Sigma_{i,j} = \sigma\rho^{|i-j|}$$

as in the other settings, but now make ρ time-varying, e.g. via

$$\rho(t) = \rho_0(1-a/2) + \rho_0(a/2)\sin\left(\frac{2\pi t}{n}\right).$$

This model would be autoregressive with lag-1 correlations oscillating between ρ_0 and $\rho_0(1-a)$, and thus have a more intuitive interpretation.

Thank you for the suggestion of this alternative setting. We agree that this definition has a more intuitive interpretation. However, we believe that the way in which ρ and ρ_0 are defined does not directly correspond to the situation we aimed to cover in Setting 4: Our goal was to mimic a situation in which the covariance structure of both observations and ensemble varies over time, with possible misspecifications of this structure in the ensemble. In the setup suggested above, the covariance structure of the observations does not change over time. Therefore, non-time-varying methods such as GCA and SSh which assume stationarity of the covariance structure are at an advantage in that they do not suffer from the (time-varying) misspecifications in the raw ensemble. An exemplary illustration is given in the attached Figure 2. Note that the rows show results for different values of a. SSh here never performs worse that ECC-Q and all methods significantly outperform ECC-Q in terms of the VS for almost all parameter values.

In order to provide a time-varying simulation with a possibly more intuitive interpretation, we have added a variant of the setting suggested above to the paper. In this new Setting 3B, we set $\Sigma^0_{i,j}(t) = \rho_0^{|i-j|}(t)$, for $i, j = 1, \dots, d$, where the correlation parameter $\rho_0(t)$ varies over iterations according to

$$\rho_0(t) = \rho_0 \cdot \left(1 - \frac{a}{2}\right) + \rho_0 \cdot \left(\frac{a}{2}\right)\sin\left(\frac{2\pi t}{n}\right)$$

for $a \in (0,1)$ for the observations, and similarly, $\Sigma_{i,j}(t) = \sigma\rho_{|i-j|}(t)$, for $i, j = 1, \dots, d$, where

$$\rho(t) = \rho \cdot \left(1 - \frac{a}{2}\right) + \rho \cdot \left(\frac{a}{2}\right)\sin\left(\frac{2\pi t}{n}\right)$$

for the ensemble predictions. Correlations in both cases are autoregressive with lag-1 and oscillated between ρ_0 and $\rho_0(1-a)$, and ρ and $\rho(1-a)$, respectively.

The description of Setting 3B was added to Section 3.3, results are discussed in Section 4.2.3, and a corresponding figure has been added to that section. Results for *additional simulation parameters for Setting 3B are provided in the Supplemental Material.*

260-264: I find this notation a bit confusing since previously the subscript/superscript 'O' was used for observations and here the subscript '0' (which is hard to discern from 'O' in the NPG font) is used to denote the fraction of zero values. The notation is also inconsistent in that in setting 3 'x' and 'y' are used to denote forecasts and observations,in contrast to the subscript/superscript 'O' for observations in the other settings.

The subscript '0' (to denote zero values) has been changed to subscript 'z'. The subscripts 'x' and 'y' have been changed to match the subscript notation in the other sections.

323 '... are identical to those of ECC-Q ...': Is this really true for ECC-S? The way it is described here, ECC-S seems to imply some level of randomization (albeit less than ECC-R), so the sampling is not the same as for ECC-Q.

Thank you for spotting this. The univariate distributions of ECC-S will indeed not be identical to those of ECC-Q. Comparing univariate results, we however found that the differences in terms of predictive performance are only very minor and were not noticeable in plots. We have modified the corresponding paragraph which now reads " Note that for ECC-S and SSh differences in the univariate forecast distributions compared to those of ECC-Q may arise from randomly sampling the quantile levels in ECC-S and due to random fluctuations due to the 10 random repetitions that were performed to account for simulation uncertainty of those methods. However, we found the effects on the univariate results to be negligible and omit ECC-S, dECC and SSh from Figure 1."

Fig. 2: I don't think that this figure is really necessary. Why is the (univariate) perfor-

mance of ECC-Q compared to the raw ensemble relevant to the comparison of multivariate postprocessing approaches?

Figure 2 and the corresponding text paragraph have been removed.

354 '... SSh never performs substantially worse ...': Why would we expect otherwise?The only drawback of SSh in the present context is the underlying assumption of time-invariance of the correlation structure, which is not a drawback in a time-invariant simulation setting. If not discussed before, this paragraph could be an opportunity to discuss this issue of time-invariant vs time-varying correlations.

We have added a short discussion to the corresponding paragraph.

360: I wonder if ECC-S gives the better results for the wrong reason here. Maybe I am misunderstanding its key idea, but to me this is essentially a compromise between ECC-Q and ECC-R. Is it possible that the small amount of randomization in ECC-S weakens the correlations, which is beneficial for $\rho > \rho_0$ and detrimental for $\rho < \rho_0$? An argument against this hypothesis is that the performance of ECC-S is not significantly different from ECC-Q when $\rho = \rho_0$. I just cannot think of any good reason why ECC-S would be better than ECC-Q. If the authors have any explanation for these results I would encourage them to include those in the discussion of the results.

Unfortunately, we are unable to provide a comprehensive explanation for these observations. As will be argued in the response to Reviewer 2 below, drawing conclusions on specific aspects of the multivariate methods or on the effect of single simulation parameters is difficult and is further impeded by a lack of well-understood multivariate evaluation methods. One potential advantage of the sampling scheme in ECC-S in comparison to the use of quantiles at fixed levels can become apparent when considering corresponding EMOS-S and EMOS-Q variants. When evaluated in terms of the VS, the random sampling in ECC-S may alleviate issues arising from over-estimated *correlations due to the use of the same quantile levels in EMOS-Q. However, it appears to not be straightforward how these observations will be effected by applying ECC, and what the effects of the 10 repetitions of each individual simulation experiments that were performed to account for simulation uncertainty, would be.*

Fig. 6: The caption should state that this is scenario B from setting 3

Fixed.

458 ’changes over iterations’: I would change this terminology and speak of ’time’ instead of ’iterations’, and be more specific about what changes/varies over time. Likewise in 464-465 I would clarify what you mean by ’structural change’

Changed as suggested.

Language and typos:

43: ... studies allow one to specifically tailor ...

Fixed.

135: sufficiently many

Fixed.

137: ’not directly straight forward’ sounds weird, I’d just say ’not straightforward’

Changed as suggested.

249: ... allows one to generate ...

Fixed.

375: ’can potentially be explained’ sounds weird, maybe better ’may be explained’

Changed as suggested.

388: I think you want to say ’In terms of ...’

Fixed.

443: Change to ’the other way round’ and ’In accordance with’

Changed as suggested.

445: Change to ’in contrast to’

Changed as suggested.

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

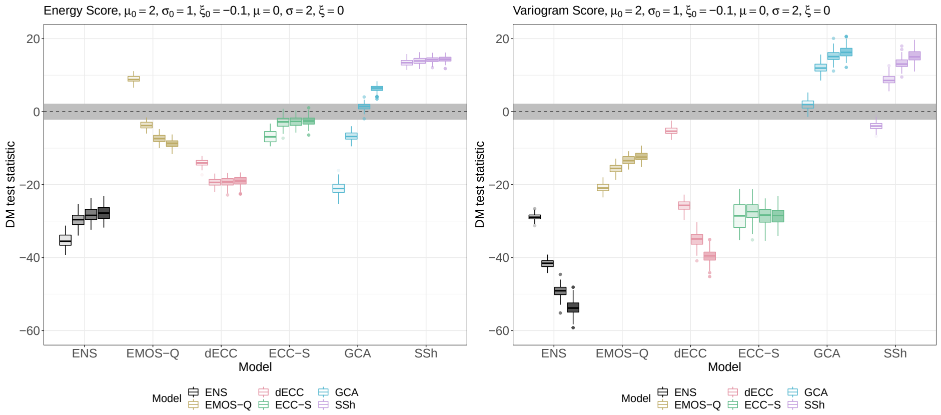

Fig. 1. Effect of ensemble size in the GEV0 setting shown as grouped boxplots of ensemble sizes $m=5,20,50,100$ (5 corresponds to lightest shade, 100 to darkest shade) for each model, for the additional Scena

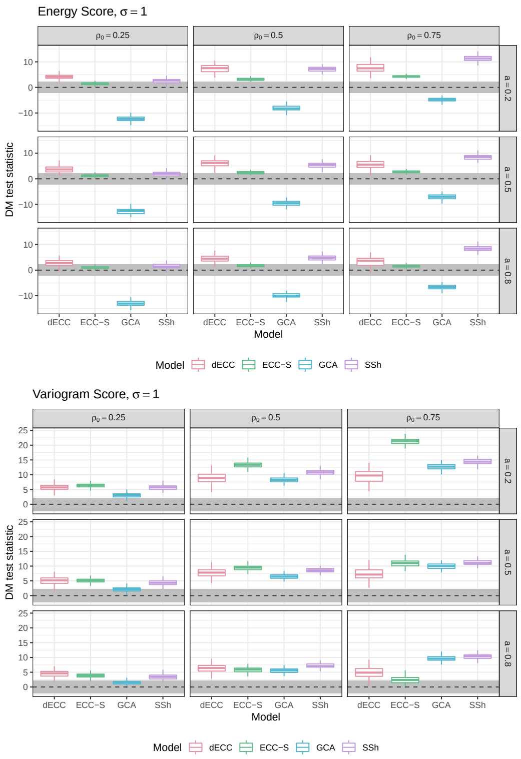

Fig. 2. Illustration similar to the figure for Setting 4 in the original main paper, but based on the adapted simulation setting suggested by Reviewer 1.