Nonlin. Processes Geophys. Discuss.,
https://doi.org/10.5194/npg-2019-62-AC3, 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

This paper present results of an intercomparison study. Different multivariate ensemble post-processing methods are compared using toy-model simulations. The focus is on empirical copula methods that are generally applied as a second post-processing step,after univariate post-processing step, in order to provide coherent multivariate structure to ensemble calibrated forecasts.

As the reviewer is the developer of one of the methods compared here, he prefers to make himself know. There exists no conflict of interest (stricto sensu) but potential cognitive biases from the reviewer side when scrutinizing the results. The paper is

clear, well structured, and well written. However, the choice regarding the selection of illustrations is not sufficiently motivated to my opinion. This choice is important because it drives the discussion and the main conclusion of the study. A 3-point argument is developed below to explain this criticism.

You claim in the conclusion (L462/463) that the 4 simulation settings aim to mimic different situations and challenges occurring in practical situations. However, the link with practical situations is sometimes weak or missing. In particular, it would be interesting to link misspecification definitions with practical examples. What sigma $>$ 1 and sigma $<$ 1 mean, and similarly what does rho $>$ rho0 and rho $<$ rho0 mean and "look like" in practice? In which situations should one expect to encounter these types of misspecification? Which type of misspecification situations are the most common in practice? Are combinations of misspecified elements more common than others (sigma $<$ 0 and rho $<$ rho0 for example)?

Once the link to the applications is clarified, illustrations could be chosen consistently. Result material is abundant, so one selection criterion could be to focus on the main misspecification encountered in practical situations. For example, you use sigma equals 1 to illustrate results in Setting 2. Does that often occur in practice? One could rather illustrate Setting 2 with sigma<1. Similarly, for Setting 3, scenario B is used for illustration purposes. It corresponds to the case where the ensemble forecasts have a heavier right tails and slightly higher point mass at zero than the observations. Does that often occur in practice? No justification for the use of this scenario as reference is provided. Scenario A (the ensemble comes from a distribution with smaller spread)or Scenario C (the observation distribution has a much heavier right tail) seem to be more likely to be faced in practice.

Thank you for this helpful comment which sparked ample discussion between the authors about how to best diagnose and transfer misspecifications of real-world ensemble prediction system to the simulation settings.

From our experience with most standard surface weather variables, one would expect that the univariate ensemble predictions are typically underdispersive ($\sigma < 1$), and exhibit a bias ($\epsilon \neq 0$). Often, biases and dispersion errors will of course vary over time, and may be vastly different for different variables or geographical locations. The choice of the correlation parameters ρ, ρ_0 naturally hinges on the chosen correlation function for the simulation setting. In practice, this would likely correspond to an over-simplification of true correlations. In an attempt to diagnose realistic correlation parameters in the context of Setting 1, we have estimated correlation parameters for 2-day ahead ECMWF ensemble predictions of 00UTC temperatures at observation stations in Germany based on the 10-year dataset used in Rasp and Lerch (2018), as follows: For a randomly selected station, we choose the 13th, 26th, 38th, 50th closest station. If there are no substantial differences in altitude, we determine the empirical correlation among the observations at those 5 stations, as well as the correlation among each ensemble member's forecasts at those stations. Next, the differences between the estimated correlation in the ensemble members and the observation are computed for all 50 members. In addition, we determine the parameters ρ, ρ_0 from the exponential correlation function model assumed in the paper by numerical optimization. The procedure described above is repeated 100 times. Figure 1 shows differences in the empirical correlation, with histograms summarizing results over all 50 members and 100 repetitions. The differences in correlation are almost identical for all close stations, suggesting that the assumption of a fixed parameter ρ and ρ_0 seems reasonable. Further, a similar conclusion can be obtained from Figure 2 which shows differences in the estimated correlation parameter of the exponential correlation model. Both figures suggest that correlations in the observations are over-estimated by the ensemble. Realistic settings thus probably relate best to simulation settings where $\rho > \rho_0$, but the values chosen in the paper (with differences of at least 0.15) appear to lead to possibly be too large differences (at least when compared to 2-day ahead temperature predictions over Germany).

In deciding on parameters for the simulation settings, we have mainly sought to cover a complete range of possible misspecifications rather than mirroring the situation in practice, in order to provide a more complete view of the performance of the multivariate post-processing methods. We have extended the discussion on the realism (or lack thereof) of the simulation settings in the Discussion, and have incorporated some of the arguments made above. Further, realistic values of the simulation parameters of Setting 1 that can be expected in practical applications are now discussed towards the end of Section 3.1. Setting 2 has been removed following the suggestion of Reviewer 1. Concerning the interpretation of the chosen scenarios in (former) Setting 3, for example Scenario B considers a situation where the forecast distribution estimates the amount of zero precipitation correctly, but otherwise the probability for obtaining a value smaller or equal to a fixed precipitation amount x computed from the forecast distribution is always smaller than the corresponding probability computed from the distribution of the observation, see the right panel in Figure 3. In combination, these two features may not occur very often in practice, since one would expect that underforecasting smaller precipitation amounts should come along with an underestimation of zero precipitation. See the left panel in Figure 3, showing fitted CDFs to a sample of observed precipitation values and corresponding forecasts of an individual ensemble member at a specific station in Germany based on real precipitation and ECMWF ensemble forecast data. Since a considerable number of scenarios with respect to the actual values of μ, σ and ξ is conceivable in practice, depending on the climatological circumstances, further investigations based on real data are required to provide additional insights. Furthermore, the interplay between the 3 GEV0 parameters is specifically complex in the sense that all 3 parameters have a joint influence on the location and dispersion properties, so that simple misspecifications in mean and variance might correspond to various (different) sets of parameter combinations. However, we agree that a more detailed investigation of theoretical and practical properties of the GEV0 distribution is a highly interesting starting point for future research.

Your conclusion points to the robustness of the SSh method and so you encourage post-processing practitioners to consider this method as their first choice. Based on the results presented in the manuscript and the ones in the supplemental material document, one could draw the opposite recommendation. First choice methods are available for different types of misspecification and so one could encourage to apply SSh only when the misspecification is unknown or difficult to identify. Specific recommendations would read:

1. If the ensemble is underdispersive and the ensemble correlation is too weak, d-ECC is the best option, both in terms of ES and VS.
2. If the ensemble is underdispersive and the ensemble correlation is too strong, ECC-S is the best option in terms of VS and one of the best options in terms of ES.
3. If the ensemble is overdispersive and the ensemble correlation is too strong, d-ECC is the best option, both in terms of ES and VS.
4. If the ensemble is overdispersive and the ensemble correlation is too weak, GCA is the best option in terms of VS, but is less performant in terms of ES. GCA proves to work well even when the overdispersiveness error characteristic is relaxed.
5. Otherwise consider using SSh, in particular when the correlation structures in theforecasts and observations are very dissimilar.

This is valid for all types of distributions (so all types of weather variables). Are these conclusions still valid in case of time varying misspecification? Verification results are missing to conclude here. Therefore, the authors are encouraged to investigate various sigmas in Setting 4 in order to collect evidences, and could draw conclusions accordingly

Thank you for these suggestions. We agree that the conclusions may benefit from a more detailed discussion of situations where the different methods show advantages or disadvantages across settings.

With the changes and additional results in the paper and Supplemental Material following the comments by Reviewer 1 (Setting 2 was removed; Setting 4 (now Setting 3A) is accompanied by another variant, Setting 3B; and additional simulations with varying numbers of ensemble members and dimensions), we believe that it is difficult to arrive at general conclusions of this form that would be valid for all types of distributions and settings. In particular, the effects of misspecifications of the individual simulation parameters may be very different in, for example, the multivariate Gaussian distribution in Setting 1 and the censored GEV variant in Setting 3 (now Setting 2). In particular for the new time-varying setting, some of the results regarding the relative performance of dECC and ECC-S, respectively, differ somewhat from the recommendations summarized above. With Setting 1 in mind, realistic parameter choices to mimic properties of real-world dataset likely represent settings where $\sigma < 1$ and $\rho > \rho_0$. As you pointed out, these settings may favor ECC-S over ECC-Q.

We have modified the discussion in Section 5 to provide more detailed suggestions and conclusions regarding the overall performance of the methods.

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

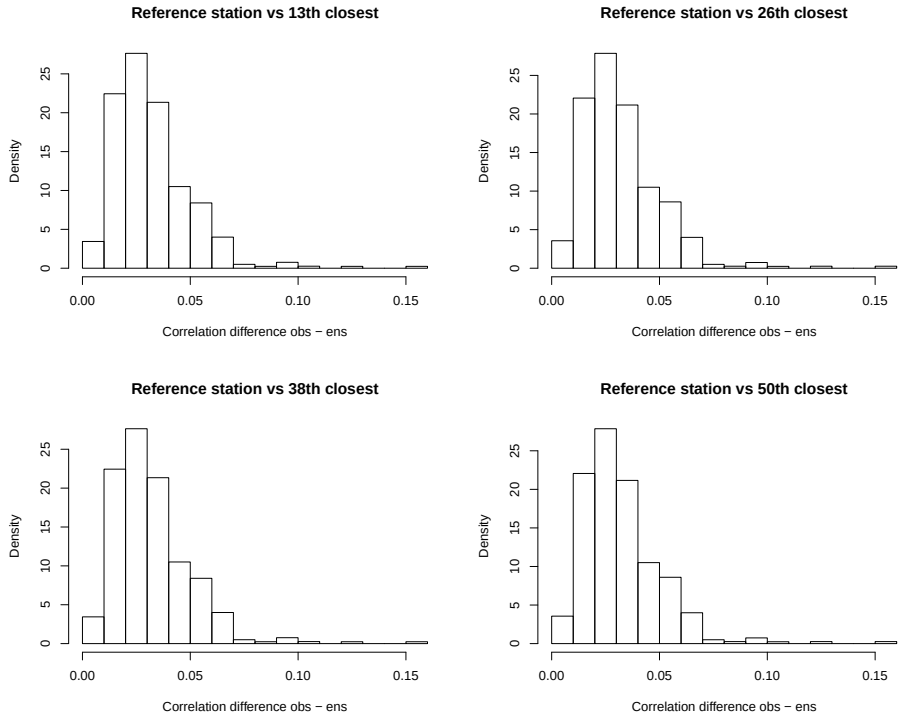

Fig. 1. Differences in empirical correlation of observations and ensemble members at a set of 5 close stations based on the dataset of Rasp and Lerch (2018).

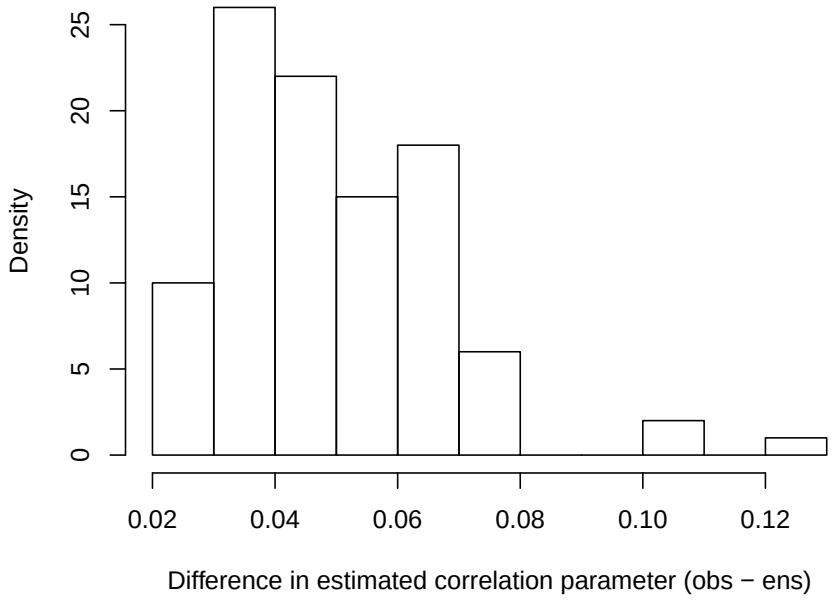

Fig. 2. Differences in estimated correlation of observations and ensemble members according to the exponential correlation function model assumed in the simulation settings at a set of 5 close stations based

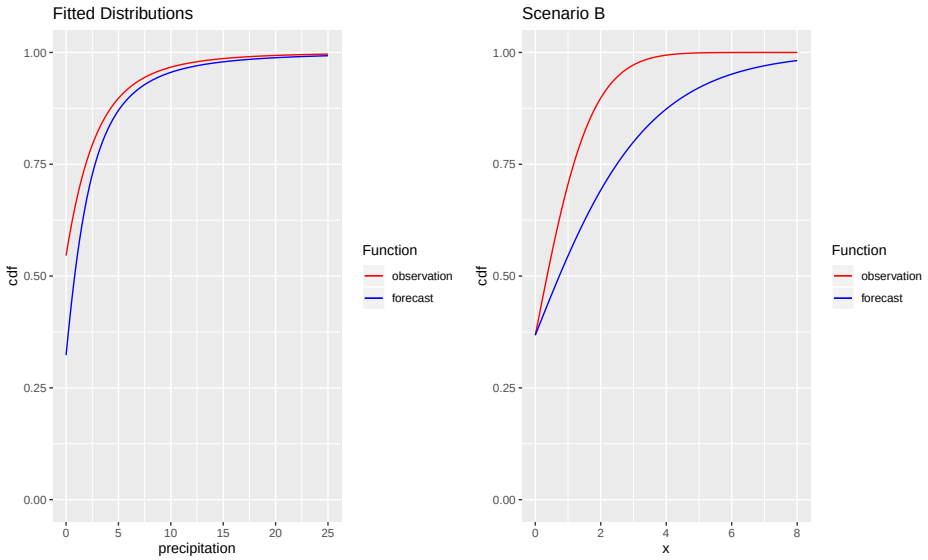

Fig. 3. Cumulative distribution functions of GEV_{0}. The fit parameters in the left panel are $\mu_{0} = -1.0698, \xi_{0} = 0.2700, \sigma_{0} = 1.9906 $ for the observation (red line) and $\mu = 0