Nonlin. Processes Geophys. Discuss.,
https://doi.org/10.5194/npg-2019-63-RC1, 2020

Nonlinear Processes in Geophysics
Discussions

Interactive comment on "Beyond Univariate Calibration: Verifying Spatial Structure in Ensembles of Forecast Fields" *by* Joshuah Jacobson et al.

Anonymous Referee #1

Received and published: 3 February 2020

General comments:

The authors study a recently developed extension of the well-known analysis rank histogram, which allows for a verification of spatial structures in ensemble forecasts. A simulation study using Gaussian random fields serves to demonstrate the connection between spatial correlation lengths and the rank histogram of the fraction of threshold exceedance (FTE). As a real world example application, the authors study statistically downscaled ensemble predictions of precipitation accumulations. The presented verification technique is interesting and could be useful to a potentially wide range of users, thanks to its relative simplicity and intuitive connection to the analysis rank histogram.

The manuscript is well written and overall easy to follow, the experiments adequately demonstrate the merits of the FTE histograms. My comments and questions mostly concern the analysis of the results. When these points have been addressed, I recommend publication of the manuscript.

1. I'm not sure that the interpretation of the FTE is always histogram quite as straightforward as you say: For the Gaussian random fields studied in section 3.3, larger correlation lengths always lead to a greater range of FTEs in the ensemble, but is this necessarily true in the general case? While I cannot, off the top of my head, name a random process that violates this assumption, I could imagine a weather prediction system which has been tuned to produce correct rain areas but systematically simulates patterns with too many small showers or too broad stratiform rain fields. In such cases, the spatial correlation structure would be misrepresented in a way that cannot be detected from the area above a given threshold alone.

2. If you follow Keller and Hense (2011) and fit a beta distribution to the histograms, why not also use their beta-score which combines a and b into a single number that characterizes over- and under-dispersion? I fail to see the advantage of considering a and b separately.

3. Related to the previous point, please explain more clearly (preferably in section 2) how you interpret the skewed cases a>b and b>a. While the skewness certainly relates to an overall over- or under-estimation of exceedance areas, maybe a more careful discussion is needed: Keller and Hense define a "beta-bias" as b-a, which is similar but not identical to the ensemble mean bias.

4. Can you comment more on the case where the margins are mis-calibrated? This is not covered by your simulation study, but the effects could potentially be a major drawback of you method. How can you be certain that the FTE really indicates errors in the spatial structure and is not determined by the erroneous marginal distribution? Could you remove the effect of the margins before calculating the FTE-histogram?

Specific comments:

page 1, l.6f: why not say exactly how the FTE is defined in the abstract? The idea is not so complicated and can be explained without equations.

page 2, l.44f: you might also cite Kapp et al. (2018) “Spatial verification of high-resolution ensemble precipitation forecasts using local wavelet spectra”, who also use wavelets but specifically only look at ensemble predictions. They also project the fields to a lower-dimensional space and then apply standard multivariate verification techniques (similar to your approach), but the interpretation of their results is not so straightforward in terms of correlation lengths.

page 4, l.93f: please explain the relationship between the correlation length and the shape of the histogram in a little more detail. I think this point is pretty central and not completely obvious.

page 4, l.97f: are you sure that cases where forecast and observations never exceed the threshold should just be discarded entirely? Shouldn’t the addition of a few correct negative cases to the verification data-set lead to a better score for the prediction system?

page 5, table 1: Consider adding a third column which states the interpretation in terms of the correlation length or the quality of the forecast. However, I’m not sure what exactly to write for a<b and a>b (see main comment 3).

page 5, l.125: If I understand correctly, the first step in this equation is only correct if F is a strictly monotonic function. What if the distribution contains a point mass, say at zero as for precipitation? Does that affect the rest of you argumentation? More generally, can you comment on how the results of your simulation study transfer to discontinuous variables like precipitation (as studied in section 4)?

page 5, l.132: Do forecast and verification really need to be correlated for this experiment? What would happen if they weren’t?

page 6, l.139: You should make it clear whether s and h are vectors or scalars, a process where C depends on a vector valued h would be anisotropic.

page 6, l.147f: I don’t understand the grammar in the sentence starting with “The multivariate ...”, is “Matern” the subject and “sets” the predicate?

page 7, l.184: Please clarify in what sense these fields are “realistic”.

page 8, Fig.2: Please make it unambiguous in the figure or caption which parameter value is a and which is b.

page 9, l.205: You should discuss the origin and interpretation of this skewness (see general comment 3). It looks like forecasts whose only shortcoming lies in the correlation length obtain a left or right skew at certain thresholds (Fig. 5). What does that mean for the interpretation of a>b and a<b in realistic situations?

page 10, Fig.5: What’s a0 here?

page 11, l220: At least for the two most extreme ratio, a and be not only tend to one but go beyond and indicate the opposite bias. How do you interpret that? Can you maybe add error bars to these plots (via bootstrapping?) to check whether some of these effects might be insignificant?

page 11, l221: While it is true that longer correlation lengths on the same grid lead to the same patterns as a decrease in domain size, are you sure that the number of available grid-points has no effect at all? Surely for very small domains it becomes increasingly hard to confidently estimate the spatial correlation structure?

page 13, l272f: why not quantify the shape of the univariate histograms via a beta-distribution as well?

page 14, l294f: Can you discuss the skewness seen in Fig.9? The difference between a and b increases with threshold but, if I understand correctly, has a different sign than seen in Fig.5 (a>b but both are smaller than 1).

page 14, l297f: I would also note that this kind of consideration could generally help you distinguish the effects of miscalibrated margins from errors in the spatial structure: If the miscalibration seen in figure 9 were caused by errors in the marginal distributions, one might expect the effect to be weakest in July were the marginal calibration was the best.

page 15, Fig.9: Some of these histograms are clearly not uni-modal. Can you comment on how your method is affected by cases where the beta-fit is likely not very good?

Interactive comment on Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2019-63, 2020.