We appreciate the time and energy these two reviewers have put into their thoughtful comments -- they have surely helped to strengthen the findings in this manuscript. Below, we address each comment with our response given in italics. We have also provided a PDF highlighting the implemented changes as a supplement to this reply.

Kind regards,
Joshuah Jacobson, William Kleiber, Michael Scheuerer, and Joseph Bellier

Reviewer 1

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 histogram is always 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.

   Our simulated fields try to simulate this effect with small correlation length corresponding to small showers and large correlation lengths corresponding to stratiform precipitation. For example, notice how the spatial patterns in the simulated binary exceedance fields of Figure 2 loosely resemble that of the newly added 5mm binary exceedance data fields in Figure 7. We agree that Gaussian random fields do not provide a comprehensive study of all possible spatial patterns that we might expect to see in practice, but for the

purposes of exploring / understanding the FTE histogram behavior in an idealized setting we believe the study is appropriate.

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.

   Thank you for this suggestion. We agree that the beta-score and beta-bias, as described in Keller and Hense (2011), are more straightforward and interpretable for characterizing over- and under-dispersion than the estimated beta-parameters alone. As such, we have updated our discussion, Table 1, and all relevant figures to focus interpretation of the FTE histogram through the lense of these metrics.

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.

   If the marginal distributions are mis-calibrated, over/under-estimation of the observed values directly translates into skewed FTE-histograms. If the marginal distributions are well-calibrated but spatial correlations are over/underestimated, skewness and cap/cup-shape of the FTE histogram are somewhat intertwined, depending on the threshold value. To see this, consider the extreme case where the ensemble forecasts are spatially uncorrelated while the observations are fully correlated. If the threshold is set to the climatological median at each grid point, the FTE for each ensemble member is close to 0.5 while the FTE for the observed field is either 0 or 1, each with equal probability. So the FTE histogram is cup-shaped with half of the cases in the lowest bin and the other half in the highest bin. If we pick a larger threshold, say the 0.95 climatological quantile, the FTE of all ensemble forecasts is close to 0.05, while the FTE of the observed field is 1 with probability 0.05 and 0 with probability 0.95. The associated FTE histogram will thus still have all cases concentrated in the lowest and the highest bin, but more (95%) cases in the lower bin, so the cup shape goes along with a skew. In reality, spatial correlations won’t be at the two extreme ends (full or no spatial dependence), but we can expect (and we see in the simulations) a similar, threshold-dependent link between cap/cup shape and skewness of FTE histograms. We have expanded our explanations regarding the link between spatial miscalibration and histogram shapes (beta parameters) such that it now includes the skewed cases.

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 your 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?

The case of mis-calibrated margins is indeed a challenge, and we added a discussion of this situation to section 2. Over- and under-forecast biases can be removed quite easily, but for dispersion errors this is not straightforward. And it is true that the effects of marginal mis-calibration and spatial mis-calibration are hard to disentangle, and that this hampers the interpretation of FTE histograms. We note though (and have added a similar statement to the manuscript) that this drawback is not specific to our approach but applies to any multivariate verification metric that is based on the idea of projecting a multivariate quantity onto a univariate one (as almost all of them do in order to condense information).

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.

  Thank you for the suggestion. We've added the definition to the abstract.

- 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.

  The suggested reference has been included and is briefly discussed.

- 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.

  We have expanded the discussion of how deficiencies in the spatial structure of the ensemble (too much or too little spatial variability) is linked to the histogram shapes (e.g., see sections 2 & 3.3.1).

- 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?

  This depends on the purpose of the metric. If the estimated beta-score of an FTE histogram were to be interpreted as a score in the traditional sense, then we agree that one should give the ensemble system credit for correctly predicting no exceedances -- in

which case artificial flattening would actually be desirable. However, we don't see the FTE method as a quantitative metric; perhaps the CRPS of the FTE is more in line with this goal (e.g., Scheuerer and Hamill, 2018, their Table 3). Instead, it is our view that the FTE histogram is best utilized as a diagnostic tool and argue that we should remove uninformative cases that flatten the histogram.

- 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).

  Table 1 and the surrounding discussion regarding interpretation of the beta-score/bias in terms of misrepresentation of spatial variability by the ensemble forecast fields have been updated, and a column with interpretations of the different shapes has been added.

- 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 your 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)?

  Our goal here is mainly to illustrate that our Gaussian process model is sufficiently general to simulate different weather variables, even if these are spatially inhomogeneous. We believe it is sufficiently general even for variables with a point mass (like precipitation). While our argument starts with the weather variable of interest and shows how it can be transformed to standard Gaussian marginal distributions, the simulation actually works the other way round: we can simulate from a Gaussian process and transform the marginal distributions to those of the desired weather variable. This is always possible, even if there is a point mass; in this case values below a certain threshold simply get mapped to zero. We have added this clarification to the manuscript.

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

  Please refer to the newly added Appendix D in which we examine a simulation experiment where the forecast and verification are explicitly uncorrelated. The short answer is that correlation is not necessary; the FTE is sensitive to miscalibration of the spatial structure, regardless of the value of correlation parameter (skill).

- 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.

  Thank you for pointing this out, our notation was not completely clean here. We fixed the problem.

- 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?

  Correct. We've updated the language for clarity.

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

  The subjective term "realistic" has been removed in favor of a direct comparison to real data fields.

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

  We've added the distinction to the caption.

- 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?

  A more rigorous discussion on the origin and interpretation of skewness has been added, and Table 1 has been updated accordingly.

- page 10, Fig.5: What's a0 here?

  a0 = 2; we've added this in the caption.

- page 11, l220: At least for the two most extreme ratios, a and b 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?

  Inclusion of the 95% confidence interval suggests that this tendency toward the opposite scoring is significant. We believe that this is due to the confounding effect of resolving ties at random, which happens more and more often (and between more and more tied FTE values) as the threshold increases, as only few ensemble member forecasts ever cross that threshold. This results in a massive distortion of histogram shapes, which still indicates non-uniformity, but makes it almost impossible to diagnose the source of this non-uniformity. We've added a discussion of this observation.

- 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?

  The number of available grid points is constant in our setup, what changes (as the correlation length of the verification field is varied) along with the `effective domain size'

is the spatial resolution. While we agree that this would have an effect on the ability to estimate the parameters of the underlying spatial correlation function, we note that such model identification is not the goal (and not required) here. For the FTE calculations, we think that the spatial resolution is not important, and the number of available grid points is only important to the extent that an insufficient number of grid points within the domain would increase the sampling variability in the calculation of the FTE values (and everything derived from them).

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

  The beta-characterization is now included with the histograms.

- 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).

  Skewness in Fig. 9 is observed primarily in months where the marginal calibration is already unreliable (i.e., Jan, April, Oct) which obfuscates interpretation of the skewness. As (now) noted in Table 1, skewness occurs at higher thresholds in combination with a U-shape, but it also occurs as a result of an under-forecast bias, and the univariate rank histograms suggest such an under-forecast bias for Jan, April, and Oct, and to a lesser degree also for July. We think that this effect is dominant here, and counteracts the spatial miscalibration-driven beta-bias effects that we would expect in conjunction with a U-shape. We have added some of this discussion to the paper.

- 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.

  Correct -- summer is really the only season in this example that permits an unambiguous conclusion about the spatial structure. This important point is given additional emphasis in our discussion (see final lines of section 4.2).

- 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?

  We believe this is just a result of sampling variability. If there is truly a multimodal shape of the histogram that is not well-represented by a beta distribution, it won't be captured by the beta-fit / beta-characterization. However, this could also be an advantage if this is merely due to sampling variability; in this case the beta distribution fit filters out some of that variability.

Reviewer 2

This manuscript discusses multivariate verification of ensemble forecasts with a focus on spatial structures. A new approach is proposed, illustrated and discussed using synthetic and precipitation datasets.

The proposed approach consists in transforming a multivariate quantity (a spatial field) into a univariate quantity. This is performed by thresholding and counting the fraction of grid points exceeding a threshold over a domain. The popular rank histogram tool is applied to the resulting ensemble and observed fractions of threshold exceedance (FTE). The method is easy to implement and is appealing because of the simplicity of both the interpretation of the derived univariate product and of the derived rank histograms.

In order to better understand FTE histograms, a toy model is used to assess the sensitivity of the method to predefined discrepancies in a controlled environment. In a real-life example context, FTE histograms are derived for precipitation forecasts and limitations of the underlying forecasting system diagnosed. The paper is well-written, well-structured, and pleasant to read. The description of the method and datasets is clear, the discussion of the results convincing.

However, the manuscript would benefit from some clarifications. First, the authors could clarify the link between the FTE histogram approach and the fraction skill score (FSS, Roberts and Lean 2008), a popular verification metric applied to deterministic forecast of precipitation fields. The first step is similar for the two methods (transformation of a real-value field into a binary field and then a focus on the fraction of events over a domain). It would be beneficial for the verification community to clearly state the link between FTE histogram and FSS.

> *Thank you for highlighting this connection; a brief discussion has been added to the introduction.*

Secondly (and more importantly), the author could discuss the impact of miscalibration of the univariate distribution on the interpretation of the FTE. It is clearly mentioned in the text that this aspect should not be disregarded. In line 86, it is stated that the FTE histogram can be used “once the marginal distribution has been checked”, and in Section 4.2 that "the possible non-uniformity of FTE histograms [...] could be due to univariate miscalibration". Perfect calibration of univariate distributions cannot be expected in reality so a discussion about the permeability of FTE histograms to univariate miscalibration would be more than useful for the interpretation of real case FTE histograms.

> *We have expanded the discussion of miscalibration of the univariate forecast distributions and its effect on the shape of the FTE histograms (both in the main text and in Table 1). When both univariate distributions and spatial correlations are miscalibrated, it is difficult to disentangle these effects. This is a consequence of projecting a multivariate quantity onto a univariate one which inevitably results in some loss of*

information. This is why we strongly advocate for considering different verification metrics - including univariate rank histograms - alongside each other. This prevents false conclusions in the worst case scenario where marginal and spatial mis-calibration cancel each other out. We note, however, that even in this case the FTE histograms (unlike other multivariate verification metrics) permit conclusions about the representation of forecast uncertainty about quantities that are meaningful in practice, so while it is not always straightforward to diagnose the exact source of mis-calibration, one can still use FTE histograms in conjunction with metrics that study e.g. regional averages or maxima to compare and rank the quality of different forecast systems.

In addition, the authors could clarify why the parameters derived from the beta distribution fitting differ from the parameters defined in the Keller and Hense 2011 paper. The list of references could also be checked and the duplicated doi information removed.

We've made an overarching transition throughout our discussion and analysis to the beta-score and beta-bias defined in Keller and Hense (2011). The duplicated doi information has also been removed from the list of references.

Beyond Univariate Calibration: Verifying Spatial Structure in Ensembles of Forecast Fields

Joshuah Jacobson[1], William Kleiber[1], Michael Scheuerer[2,3], and Joseph Bellier[2,3]

[1]Department of Applied Mathematics, University of Colorado Boulder
[2]Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder
[3]Physical Sciences Laboratory, National Oceanic and Atmospheric Administration, Boulder, Colorado

Correspondence: Joshuah Jacobson (Josh.Jacobson@Colorado.edu)

Abstract. Most available verification metrics for ensemble forecasts focus on univariate quantities. That is, they assess whether the ensemble provides an adequate representation of the forecast uncertainty about the quantity of interest at a particular location and time. For spatially-indexed ensemble forecasts, however, it is also important that forecast fields reproduce the spatial structure of the observed field, and represent the uncertainty about spatial properties such as the size of the area for which heavy precipitation, high winds, critical fire weather conditions, etc. are expected. In this article we study the properties of the fraction of threshold exceedance (FTE) histogram, a new diagnostic tool designed for spatially-indexed ensemble forecast fields. ~~The metric is based on a level-crossing statistic that we term~~ Defined as the fraction of ~~threshold exceedance (FTE), and~~ grid points where a prescribed threshold is exceeded, the FTE is calculated for the verification field, and separately for each ensemble member. ~~The FTE~~ It yields a projection of a – possibly high-dimensional – multivariate quantity onto a univariate quantity that can be studied with standard tools like verification rank histograms. This projection is appealing since it reflects a spatial property that is intuitive and directly relevant in applications, though it is not obvious whether the FTE is sufficiently sensitive to misrepresentation of spatial structure in the ensemble. In a comprehensive simulation study we find that departures from uniformity of ~~these so called~~ the FTE histograms can ~~be~~ indeed be related to forecast ensembles with biased spatial variability, and that these histograms detect shortcomings in the spatial structure of ensemble forecast fields that are not obvious by eye. For demonstration, FTE histograms are applied in the context of spatially downscaled ensemble precipitation forecast fields from NOAA's Global Ensemble Forecast System.

1 Introduction

Ensemble prediction systems like the ECMWF ensemble (Buizza et al., 2007) or NOAA's Global Ensemble Forecast System (GEFS; Zhou et al., 2017) are now state of the art in operational meteorological forecasting at weather prediction centers worldwide. One of the goals of ensemble forecasting is the representation of uncertainty about the state of the atmosphere at a future time (Toth and Kalnay, 1993; Leutbecher and Palmer, 2008), and verification metrics are required that can assess to

what extent this goal is achieved. For univariate quantities, i.e. if forecasts are studied separately for each location and each forecast lead time, diagnostic tools like verification rank histograms (Anderson, 1996; Hamill, 2001) or reliability diagrams (Murphy and Winkler, 1977) can be used to check whether ensemble forecasts are calibrated, i.e. statistically consistent with the values that materialize.

When entire forecast fields are considered, aspects beyond univariate calibration are important. For example, ensembles that yield reliable probabilistic forecasts at each location may still over- or under-forecast regional minima/maxima if their members exhibit an inaccurate spatial structure (e.g., Feldmann et al., 2015, their Fig. 6). For weather variables like precipitation, which are used as inputs to hydrological forecast models, it is crucial that accumulations over space and time (and the associated uncertainty) are predicted accurately, and this again requires an adequate representation of spatial structure and temporal persistence of precipitation by the ensemble.

There is an added difficultly for forecasters in that misrepresentation of the spatial structure of weather variables by ensemble forecast fields may not be discernible by eye. For example, consider the simulated fields in Fig. 1: perhaps one of these forecast fields has a clearly different spatial correlation length than the verification, but we suspect that even the sharp-eyed reader cannot distinguish between the remaining fields with confidence. Even if the differences are obvious, a quantitative verification metric is required to objectively compare different forecast systems or methodologies.

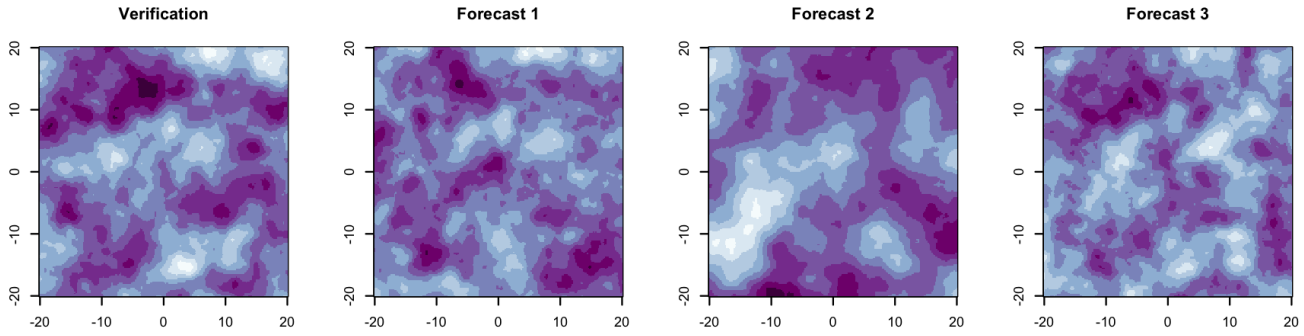

Figure 1. Simulated verification field and three associated forecast fields (arbitrary color scale) in which the spatial correlation length is either the same as for the verification, 10% miscalibrated, or 50% miscalibrated. Can you tell which is correct?

Several multivariate generalizations of verification rank histograms such as minimum spanning tree histograms (Smith and Hansen, 2004; Wilks, 2004), multivariate rank histograms (Gneiting et al., 2008), average-rank and band-depth rank histograms (Thorarinsdottir et al., 2016), and copula probability integral transform histograms (Ziegel and Gneiting, 2014) have been proposed and allow one to assess different aspects of multivariate calibration. They are all based on different projections of the multivariate quantity of interest onto a univariate quantity that can then be studied using standard verification rank histograms. Unfortunately, most of these projections do not allow an intuitive understanding of exactly what multivariate aspect is being assessed, and none are tailored to the special case where the multivariate quantity of interest is a spatial field. ~~A recent paper by Buschow et al. (2019) proposes~~ Two recent papers (Buschow et al., 2019; Buschow and Friederichs, 2020) propose a wavelet-

based verification approach in which wavelet transformations of forecast and observed fields are performed to characterize and compare the fields' texture. The authors demonstrate that this approach is able to detect differences in the spatial correlation length similar to those shown in Fig. ~~1.~~ 1. Kapp et al. (2018) define a skill score based on wavelet spectra and study the score differences between a randomly selected ensemble member and the verification field in order to detect possible deficiencies in the texture of the forecast fields. Our goal is similar, but the approach studied here ~~is more in line with~~ follows the idea of defining a projection from the multivariate quantity (here: a spatial field) to a univariate quantity that can be analyzed via verification rank histograms. ~~We are also more focused~~ Our main focus is on the probabilistic nature of the forecasts~~. That~~, that is, we ~~are studying whether the forecasts adequately represent spatial variability *as an ensemble*~~want to test whether the ensemble adequately represents the *uncertainty* about spatial quantities.

The projection underlying the verification metric studied here is based on threshold exceedances of the forecast and observation fields. This binarization of continuous weather variables is common in spatial forecast verification (see Gilleland et al., 2009) as it allows one to study, for example, low, intermediate, and high precipitation amounts separately. In ~~a recent paper, Scheuerer and Hamill (2018) calculate the fractions of threshold exceedance (FTE)~~ the context of deterministic forecast verification, Roberts and Lean (2008) define the fractions skill score (FSS) based on the fraction of threshold exceedances (FTEs) within a certain neighborhood of every grid point, and use it to examine at which spatial scale the forecast FTEs become skillful. Scheuerer and Hamill (2018) use a similar concept to study whether an ensemble of forecast fields adequately represents spatial forecast uncertainty. They calculate the FTE for all ensemble members and the verifying observation field, and study verification rank histograms of the resulting univariate quantity in order to diagnose the advantages and limitations of different statistical methods to generate high-resolution ensemble precipitation forecast fields based on lower resolution NWP model output. The FTE is an interpretable quantity that is highly relevant in applications where the fraction of the forecast domain for which severe weather conditions are expected (e.g., heavy rain, extreme wind speeds, etc.) may be of interest. However, it is not obvious whether FTE histograms are sufficiently sensitive to misrepresentation of the spatial structure by the ensemble, and the goal of the present paper is to investigate this discrimination ability in detail.

In section 2, we describe the calculation of the FTE and the construction of the FTE histogram in detail. In section 3, a simulation study is designed and implemented that allows us to analyze the discrimination capability of the FTE histograms with regard to spatial structures. In section 4, we demonstrate the utility of FTE histograms in the context of spatially downscaled ensemble precipitation forecast fields from NOAA's Global Ensemble Forecast System. A discussion and concluding remarks are given in section 5.

2 The fraction of threshold exceedance metric

Let $Z(s)$ be a scalar field on a domain $s \in D$. Here, we describe a strategy of studying exceedances of Z at various thresholds. That is, we focus interest in statistics based on $\mathbf{1}_{\{Z(s)>\tau\}}$ for a given threshold $\tau \in \mathbb{R}$. In the domain D, we define the fraction

of threshold exceedance (FTE) as the fraction of all points at which τ is exceeded. Specifically, let

$$\begin{aligned}\mathrm{FTE}(Z,\tau) &= \frac{1}{|D|}\int_D \mathbf{1}_{\{Z(s)>\tau\}}(s)ds \\ &= \frac{1}{n}\sum_{j=1}^{n}\mathbf{1}_{\{Z(s)>\tau\}}(s_j) \end{aligned} \quad (1)$$

where the first equality represents the idealized continuous spatial process definition, while the second reflects the discrete nature of spatial sampling in an operational probabilistic forecasting context with $D = \{s_1, \ldots, s_n\}$. The resulting univariate quantity can be evaluated by common univariate verification metrics (Scheuerer and Hamill, 2018).

Suppose we have a k-member ensemble $Z_1(s), \ldots, Z_k(s)$ and associated verification field $Z_0(s)$ (e.g., observation or analysis) all on D; let $\pi = \{\mathrm{FTE}(Z_0,\tau), \ldots, \mathrm{FTE}(Z_k,\tau)\}$. Note that π depends on the threshold, but for ease of exposition we do not include this dependence in notation. We call r the rank of the verification FTE relative to the set of verification and ensemble forecast FTEs, or the rank of $\mathrm{FTE}(Z_0,\tau)$ in π. There are three cases of interest when computing r: (1) no ties exist in π, (2) ties exist among a subset of π that includes $\mathrm{FTE}(Z_0,\tau)$, or (3) there is only one unique value in π. In the first case no special action is required, and in the second case ties in rank are simply broken uniformly at random. The third case arises when all ensemble members have the exact same FTE as the verification, as may occur ~~e.g.~~, for example, when the precipitation amount reported by the verification and predicted by all ensemble members is below the threshold τ everywhere in D. Instances of this case are completely uninformative for the purpose of diagnosing miscalibration and can be discarded.

Gathering ranks over N instances of forecast/verification pairs, $r_1, \ldots, r_N$, a natural way to communicate the FTE rank behavior is through a histogram ~~that we term the~~ (termed *FTE histogram* by Scheuerer and Hamill (2018)) over the $k+1$ possible ranks. ~~In construction , the FTE histogram is similar to the (univariate)~~ Its construction is akin to that of the univariate verification rank histogram discussed in Anderson (1996) and Hamill (2001)~~. That rank histogram, however,~~ , but the latter only evaluates the marginal distribution ~~, or point-wise accuracy~~ of the ensemble~~, and may be considered a first step in the verification procedure. Once the marginal distribution has been checked, the FTE histogram can be used to measure spatial calibration of the ensemble~~. ~~An ensemble that is spatially calibrated will exhibit the same effective correlation length as the verification (see Fig.1).As in the univariate case, flat FTE histograms are a necessary (but not sufficient!) condition for reliability as they indicate that the verification and ensemble are indistinguishable with regard to the particular aspect of the forecast fields (here: fraction of threshold exceedance) assessed by this metric. Assuming that calibration of the marginal distributions has already been confirmed, non-flat FTE histograms can be related to systematic differences in the spatial structure of the verification and the ensemble forecast fields. If the correlation length of the ensemble fields is too large, a~~ The FTE histogram behaves similar to the univariate verification rank histogram under marginal miscalibration in that overpopulated low (high) bins are an indication of an over-forecast (under-forecast) bias, and a $\cup$–shaped ($\cap$~~–shaped histogram can be expected; that is~~ –shaped) histogram is an indication of an under-dispersed (over-dispersed) ensemble. However, it is also sensitive to misrepresentation of spatial correlations by the ensemble forecast fields. To see this, consider first the extreme case where the forecast fields are spatially uncorrelated (i.e., ~~verification ranks are centrally overpopulated because the larger correlation length of the ensemble makes it more likely that if one grid point is above (below) the threshold, many of the grid points in its vicinity are also above~~

(below)the threshold, and thus the ensemble FTEs often take on very low or very high ranks . Conversely spatial white noise) while the verification fields have maximal spatial correlations. In this setup, if τ is equal to the climatological median of the marginal distributions at each grid point, the FTE for each ensemble member is close to 0.5 while the FTE for the verification field is either 0 or 1, with equal probability. The associated FTE histogram is $\cup$–shaped with half of the cases in the lowest bin and the other half in the highest bin. If τ is equal to the $95th$ climatological percentile, the FTE of each ensemble member is close to 0.05 and the FTE of the verification field is 0 with probability 0.95 and 1 with probability 0.05. The associated FTE histogram is $\cup$–shaped *and* skewed, with 95% of all cases in the lowest bin and 5% of all cases in the highest bin. For τ equal to the $5th$ climatological percentile, the skewness is in the other direction with 5% (95%) of all cases in the lowest (highest) bin. In a more realistic situation, where both forecast end verification fields are spatially correlated but the spatial correlations of the forecast fields are too weak (i.e., they exhibit too much spatial variability), ~~a~~ we can still expect to see a somewhat $\cup$–shaped ~~histogram is taken as sign that the correlation length of the ensemble is too small, resulting in verification ranks becoming excessively populated to the left or right as ensemble FTEs consistently take on intermediate ranks. We note that cases where π has only one unique value~~ FTE histogram since the verification FTE values are more likely to assume extreme ranks than the ensemble FTE values. For large values of τ, the lower bins will be more populated, for small values of τ, the higher bins will be more populated. Conversely, if the spatial correlations of the forecast fields are too strong, the verification ranks will over-populate the central bins, slightly shifted upward or downward from the center depending on τ. An ensemble that is marginally and spatially calibrated (i.e., ~~all forecast and verificationFTEs are the same)are completely uninformative and can be discarded in calculating the FTE histogram in order to avoid artificial uniformity~~the strength of spatial correlations within each forecast field matches that of the verification) will result in a flat FTE histogram.

If the marginal forecast distributions are miscalibrated, the resulting effects on the rank of the verification FTE are superimposed on those caused by misrepresentation of spatial correlations. This complicates interpretation because it is often impossible to disentangle the different sources of miscalibration (this loss of information is an inevitable consequence of projecting a multivariate quantity onto a univariate one), and it can even happen that different effects cancel each other out. For example, ensemble forecast fields which are both under-dispersive and have too strong spatial correlations may result in flat FTE histograms. This serves as a reminder that – as in the univariate case – a flat histogram is a necessary but not a sufficient condition for probabilistic calibration. It simply indicates that the verification and the ensemble are indistinguishable with regard to the particular aspect of the forecast fields (here: exceedance of a prespecified threshold) assessed by this metric. Systematic over- or under-forecast biases can be accounted for by using different (depending on the respective climatology) threshold values τ for the forecast and verification fields. We are not aware of an equally straightforward way to account for dispersion errors, so we encourage users to always check the marginal forecast distributions first, and then study FTE histograms for different thresholds, possibly in conjunction with other multivariate verification metrics in order to obtain a comprehensive picture of the multivariate properties of the ensemble forecasts.

While the FTE histogram is ~~visually intuitive~~a useful visual diagnostic tool, a quantitative measure for studying departures from uniformity is desirable. Akin to Keller and Hense (2011), we ~~summarize the FTE histogram ,~~ fit a beta distribution to the histogram values (transformed to the unit interval~~, with two parameters from a beta distribution . However, as~~) and characterize

the histogram shape based on the β-*score* and β-*bias*, respectively defined as

$$\beta_S = 1 - \sqrt{\frac{1}{a \cdot b}}, \quad \beta_B = b - a, \tag{2}$$

where a and b are the two distribution parameters. Since histogram values only occur at discrete points in $[0,1]$, parameter estimation methods will incur some bias due to the lack of data on the interior of adjacent ranks. Thus, we stochastically disaggregate the (transformed) ranks $r_1, \ldots, r_N$ to continuous values in $[0,1]$ ~~;~~ (see Appendix A for details~~. We then~~) and fit a beta distribution ~~to the disaggregated ranks by~~ via maximum likelihood. ~~This provides~~ Together, the β-score and β-bias provide a pair of succinct descriptive statistics ~~in the form of the estimated shape parameters a and b, which respectively describe the behavior of the left and right sides of the histogram~~which communicate the visual characteristics of the histogram, and therefore the ensemble's calibration properties. In the ideal case, ~~a and b~~ β_S and β_B are both exactly ~~one~~zero, indicating that the FTE histogram is perfectly uniform. In practice, these ~~parameters~~ metrics are never exactly ~~one~~zero. The resulting set of possible ~~parameter combinations and corresponding departures from uniformity~~ deviations and broad interpretations of the corresponding histogram shapes are outlined in Table 1. With ~~parameters a and b~~ the β-score and β-bias, we have ~~obtained~~ an easily interpreted measure of spatial forecast calibration.

Table 1. ~~Possible departures from~~ Characterization of FTE histogram ~~uniformity, as characterized by beta parameter relationships~~shapes via β-score and β-bias and their interpretation with regard to potential deficiencies of the ensemble forecast fields.

~~Parameter~~ Histogram ~~Relationship~~	Parameters	Score & Bias	Interpretation
Uniform	$a = b = 1$	~~Uniform $a, b < 1$~~ $\beta_S = \beta_B = 0$	Ensemble FTEs consistent with verifica
∪–shaped ~~$a, b > 1$~~	$a, b < 1$	$\beta_S < 0$	Under-dispersed marginal distributions OR excessi
∩–shaped	$a, b > 1$	$\beta_S > 0$	Over-dispersed marginal distributions OR insufficie
Right-skewed	$a < b$	~~Right-skewed~~ $\beta_B > 0$	Over-forecast bias OR excessive spatial variability
Left-skewed	$a > b$	~~Left-skewed~~ $\beta_B < 0$	Under-forecast bias OR insufficient spatial variabili

Skewness is exaggerated by high thresholds; see text for more detail.

In summary, the FTE metric is composed of three steps: (1) calculate the FTE of each verification and ensemble forecast field, (2) construct an FTE histogram over available instances of forecast and verification times, and (3) ~~fit beta parameters to the~~ derive the β-score and β-bias from the stochastically disaggregated FTE histogram ~~by maximum likelihood~~ to characterize departure from uniformity.

3 Simulation study

In this section we consider an extensive simulation study to assess the ability of the proposed FTE histogram ~~in diagnosing mismatches between the correlation length of~~ to diagnose deficiencies in the ~~forecast fields and that of the verification fields. It~~

~~is not straightforward to define what a "correlation length" is in general for~~ representation of spatial variability by the ensemble forecast fields. Our simulations will be based on multivariate Gaussian processes where the notion of "spatial variability" can be quantified in terms of a correlation length parameter. The various meteorological quantities of interest such as precipitation and wind speeds ~~, especially given possible heterogeneity and spatial nonstationarity~~ can be quite heterogeneous and spatially nonstationary over the study domain. However, ~~the effect of using a threshold exceedance helps mitigate this problem. Consider~~ since we study the spatial structure of threshold exceedances, a suitable choice of thresholds can mitigate these effects to a degree that multivariate, stationary Gaussian processes can be viewed as a sufficiently flexible model for simulating realistic spatial fields. To see this, consider a strictly positive and continuous variable $Z(s)$ at two spatial locations $s = s_1, s_2$ with possibly unequal continuous cumulative distribution functions F_1 and F_2, respectively. Rather than considering a spatially-constant threshold such as 10 m/s for wind gusts, ~~it is natural to consider~~ we can use a location-dependent threshold, say the 90% climatological quantiles $q(s_1)$ and $q(s_2)$ representing local characteristics. Then both quantities $\mathbf{1}_{\{Z(s_i)>q(s_i)\}}$, $i = 1, 2$, are identically distributed Bernoulli(0.1) random variables~~, $i = 1, 2$~~. Exploiting a standard Gaussian probability integral transformation method, we note that $\Phi^{-1}(F(Z(s_i)))$ is a standard normal random variable, where Φ is the cumulative distribution function of a standard normal. Thus, the original probability of threshold exceedance can be written

$$P(Z(s_i) > q(s_i)) = P(F(Z(s_i)) > F(q(s_i))) = P(\Phi^{-1}(F(Z(s_i))) > \Phi^{-1}(F(q(s_i)))) = P(X > \Phi^{-1}(0.9)) \tag{3}$$

where X is a standard normal. Thus, we have shown that a field of random variables with continuous, possibly distinct local probability ~~distribution~~ distributions can be transformed to ~~Gaussian, and if we use~~ standard Gaussian marginal distributions, and using local quantiles as the threshold ~~then this is~~ is then equivalent to a spatially-constant threshold on the transformed variables. For weather variables with discrete-continuous marginal distributions (e.g., precipitation), this direction of the transformation is not quite as straightforward. Conversely, however, simulated fields from Gaussian processes can always be transformed to any desired marginal distributions (including discrete-continuous ones). In our ensuing simulations studies we therefore consider stationary spatial Gaussian processes as representing forecast and verification fields.

The main technical difficulty in setting up the simulation study is in generating multiple, stationary Gaussian random fields that have different correlation lengths while being correlated with each other. That is, we would like to generate $Z_0(s)$ and $Z_1(s)$ in such a way that $\mathrm{Cov}(Z_0(s), Z_1(s)) > 0$ (representing that the forecast field is correlated with the verification field) and where Z_0 and Z_1 have possibly distinct correlation lengths (representing that the forecast field is spatially miscalibrated). A natural approach is to use multivariate random field models.

3.1 Multivariate Gaussian processes

We call a vector of processes $(Z_0(s), Z_1(s), \ldots, Z_k(s))$ a multivariate Gaussian process if its finite-dimensional distributions are multivariate normal. We focus on second-order stationary mean zero multivariate Gaussian processes in that $E(Z_i(s)) = 0$ for all $i = 0, \ldots, k$ and $s \in D$. Stationarity implies that the stochastic process is characterized by

$$C_{ij}(h) = \mathrm{Cov}(Z_i(s+h), Z_j(s)), \quad \text{for all } h \text{ such that } s + h \in D, \tag{4}$$

which are called covariance functions for $i = j$ and cross-covariance functions for $i \neq j$. Not all choices of functions C_{ij} will result in a valid model, in particular we require that the matrix of functions $\mathbf{C}(h) = (C_{ij}(h))_{i,j=0}^{k}$ be a nonnegative definite matrix function, the technical definition of which can be found in Genton and Kleiber (2015).

There are many models for multivariate processes (Genton and Kleiber, 2015), and here we exploit a particular class called the multivariate Matérn (Gneiting et al., 2010; Apanasovich et al., 2012). We rely on the popular Matérn correlation function

$$M(hd|\nu, a) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{h}{a} \frac{d}{a} \right)^{\nu} K_{\nu} \left(\frac{h}{a} \frac{d}{a} \right) \tag{5}$$

where Γ is the gamma function, K_ν is the modified Bessel function of the second kind of order ν and d is a non-negative scalar. Parameters have interpretations as a ~~smoothing~~ smoothness (ν), and ~~correlation length or spatial range~~ spatial range or correlation length (a). The multivariate Matérn ~~sets~~ correlation function is defined as

$$C_{ii}(h) = \sigma_i^2 M(\|h\||\nu_i, a_i), \qquad \text{for } i = 0, \ldots, k, \tag{6}$$

and

$$C_{ij}(h) = C_{ji}(h) = \rho_{ij}\sigma_i\sigma_j M(\|h\||\nu_{ij}, a_{ij}), \qquad \text{for } 0 \leq i \neq j \leq k. \tag{7}$$

where $\|\cdot\|$ is the Euclidean norm. In this latter equation, ~~$\rho_{ij} \in [0, 1]$~~ $\rho_{ij} \in [-1, 1]$ is the co-located cross-correlation coefficient. Interpretation of the cross-covariance parameters requires spectral techniques (Kleiber, 2017).

3.2 Simulation setup

Simultaneously simulating the verification field $Z_0(s)$ and all forecast fields $Z_1(s), \ldots, Z_k(s)$ is difficult due to the high-dimensional joint covariance matrix. Instead, we approach simulations by jointly simulating the verification fields $Z_0(s)$ and the (scaled) ensemble mean field, $Z_M(s)$ from a bivariate Matérn model. We then perturb the mean field with independent univariate Gaussian random fields to generate an 11-member ensemble, $k = 11$.

The simulation setup follows a series of steps:

1. Generate Z_0 and Z_M, the verification and (scaled) ensemble mean as a mean zero bivariate Gaussian random field with multivariate Matérn ~~range~~ correlation length parameters a_0, a_M, and $a_{0M} = \sqrt{a_0 a_M}$, smoothness parameters $\nu_0 = \nu_{0M} = \nu_M = 1.5$, and co-located correlation coefficient $\rho_{0M} = \omega = 0.8$.

2. Generate 11 independent mean zero Gaussian random fields $W_1(s), \ldots, W_{11}(s)$ with Matérn covariance having ~~range~~ correlation length $a = a_M$ and smoothness $\nu = \nu_M = 1.5$.

3. The ensemble member fields $Z_1(s), \ldots, Z_{11}(s)$ are constructed as

$$Z_i(s) = \omega Z_M(s) + \sqrt{1 - \omega^2} W_i(s), \qquad i = 1, \ldots, 11. \tag{8}$$

The third step implies that each field in the ensemble is a Gaussian process with mean zero, variance one, ~~range~~ correlation length a_M, smoothness ν_M, and univariate "forecast skill" controlled by the parameter ω (see Appendix B). Note that by choosing the co-located correlation coefficient $\rho_{0M} = \omega$, the correlation between the verification and each ensemble member is ω^2, the same as the correlation between ensemble members themselves. That is, $\text{Cov}[Z_i, Z_j] = \omega^2$ for $i, j = 0 \ldots, 11$ when $i \neq j$ (derivation in Appendix C), and thus the ensemble forecasts are calibrated in the univariate sense.

In this study, fields were constructed on a square grid ~~[−20, 20]~~ over the domain $[-20, 20] \times [-20, 20]$ with resolution 0.2. ~~The simulation above was repeated five-thousand~~ Verification-ensemble samples were collected by repeating the simulation above 5000 times for each combination of

$a_0 \in \{1, 1.5, \ldots, 3.5, 4\},$

$a_M \in \{0.5a_0, 0.6a_0, \ldots, 1.4a_0, 1.5a_0\},$

~~for~~

$a_0 \in \{1, 1.5, \ldots, 3.5, 4\}, \quad a_M \in \{0.5a_0, 0.6a_0, \ldots, 1.4a_0, 1.5a_0\},$

resulting in a total of ~~seventy-seven parameters sets investigated. For~~ 77 experiments. Note that in practice, each sample corresponds to a date for which forecasts have been issued and verifying observations are available, meaning the sample size is governed by the time period for which the verification is performed. For each experiment, FTE histograms were constructed from the 5000 samples using each ~~parameter set, we constructed FTE histograms from the five-thousand samples based on each~~ of $\tau \in \{0, 0.5, \ldots, 3.5, 4\}$. That is, for a given a_0 and a_M we analyzed nine FTE histograms, for a total of 693 histograms across all ~~parameter sets~~experiments.

3.3 Simulation analysis

The question of primary interest in this analysis is whether the FTE histogram accurately identifies miscalibration of ensemble correlation lengths.

3.3.1 Illustrative examples of FTE histograms

First, we study the discrimination ability of the FTE histogram in something of an exaggerated setting, where the miscalibration is obvious. We choose the ~~mean~~ median of the marginal distribution as the threshold (i.e., $\tau = 0$) and a verification correlation length of ~~2 which we found to produce "realistic" fields on this grid.~~2. On this grid, binary fields produced in this way appear qualitatively similar to the binary precipitation fields analyzed later in this manuscript (see Fig. 7). The correlation length ratio is the ratio of the ensemble correlation length to that of the verification field. We study ensembles with too small of a correlation length using ratio 0.5 (Fig. 2, row A), correct correlation length using ratio 1.0 (Fig. 2, row B), and too large of a correlation length using ratio 1.5 (Fig. 2, row C). Corresponding FTE histograms are then constructed ~~for these three sets of verification and ensemble fields with~~ with respect to these three ratios using 5000 verification-ensemble samples in each ~~set~~case. This revealing example is depicted in Fig. 2 and behaves as described in Table 1, where the FTE histogram takes a

∪-shape (∩-shape) when the ensemble correlation length is too small ~~, a ∩-shape when the correlation length is too large, and is~~ (large), indicating excessive (insufficient) spatial variability. As desired, the FTE histogram is approximately flat when the ensemble fields have the same correlation length as the verification field.

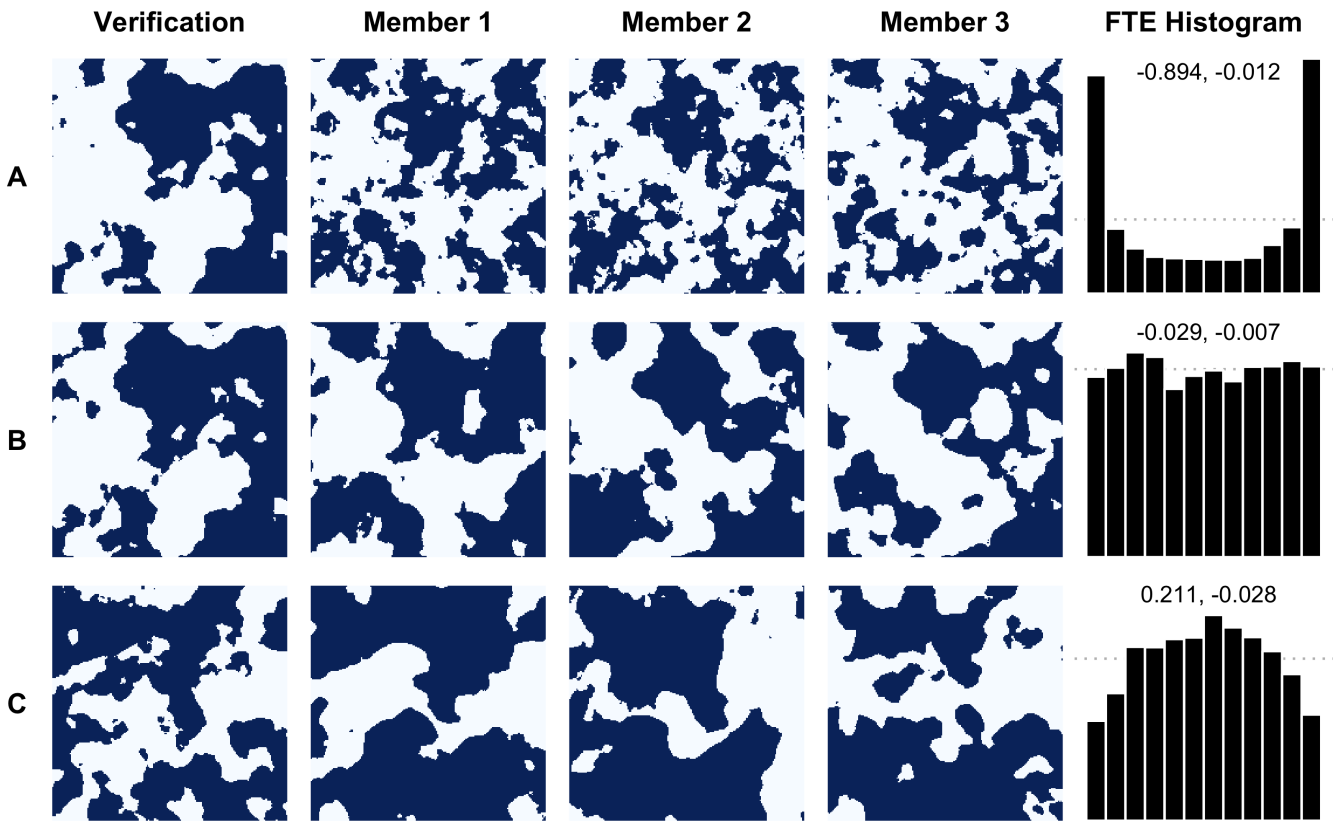

Figure 2. Example binary exceedance verification field and a subset of ensemble fields with representative FTE histogram for threshold $\tau = 0$ found using 5000 samples. Dark blue regions indicate threshold exceedance. All verification fields have correlation length $a_0 = 2$ and ensemble fields have correlation length $a_M = 1, 2, 3$ in rows A, B, and C respectively. FTE histograms are density histograms with ~~a~~ dotted ~~grey~~ line $y = 1$ and ~~estimated beta distribution parameters~~ corresponding β-score (left) and β-bias (right) annotated.

While the FTE histogram is able to correctly identify the obvious miscalibration of the ensemble for the scenario in Fig. 2, one could likely draw the same conclusions by visual inspection and would not use the FTE histogram for these fields in practice. However, ensemble forecast models are not generally so grossly miscalibrated; though a true correlation length ratio does not exist in reality, the theoretical ratio will often be much closer to unity. Therefore, the true utility of the FTE histogram is realized when the miscalibration is not so visually obvious. This more realistic example is illustrated in Fig. 3 where the above experiment is repeated using different correlation length ratios. In row A, the ensembles have ratio 0.9 and the resulting FTE histogram is still noticeably ∪-shaped. The ratio in row B is 1.0 which yields a flat FTE histogram. In row C, the ratio is

1.1 and the FTE histogram is noticeably ∩-shaped. Again, these results are consistent with Table 1, and we conclude that the FTE ~~metric~~ histogram maintains accurate discrimination ability even when ensemble members are only slightly miscalibrated.

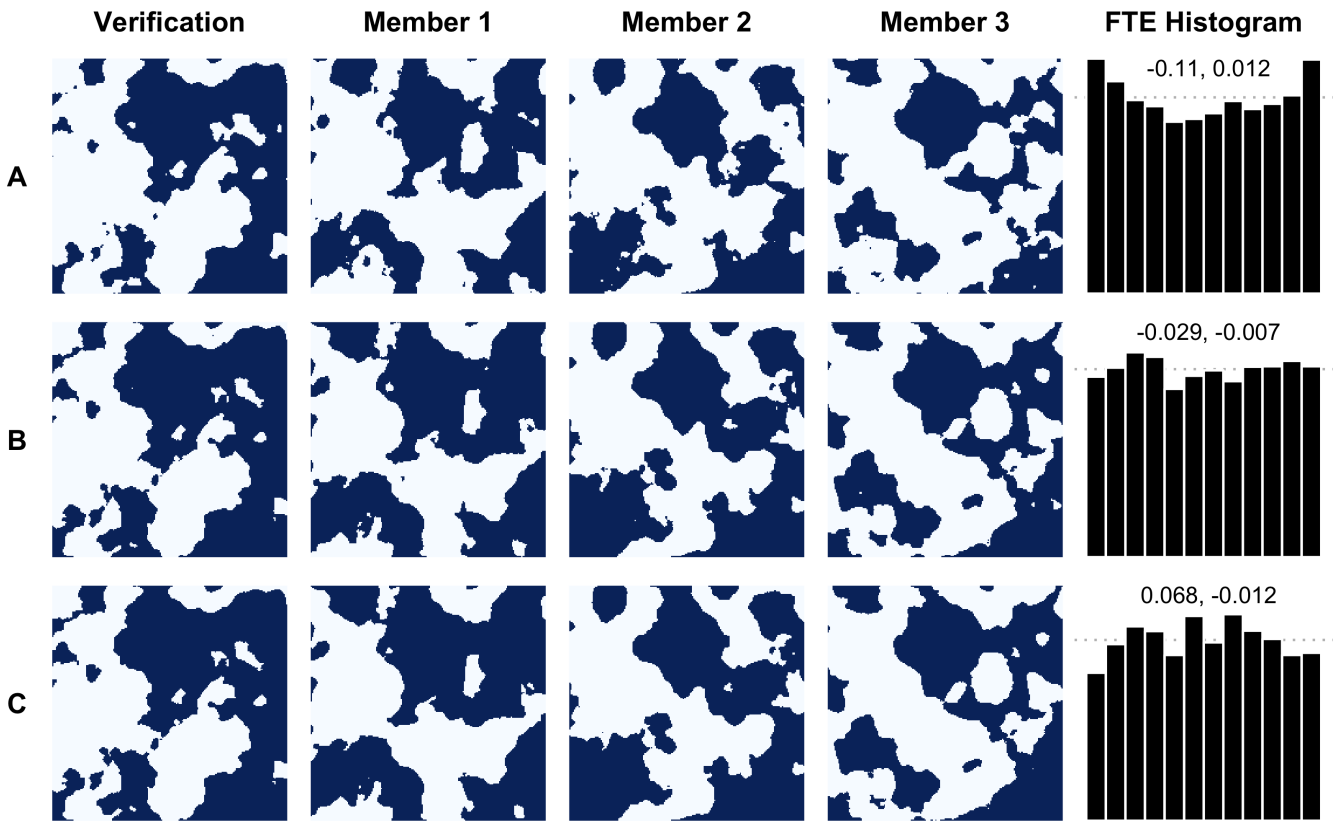

Figure 3. ~~Same as~~ As Fig. 2, but ensemble fields have ~~ranges~~ correlation length $a_M = 1.8, 2, 2.2$ in rows A, B, and C respectively.

Of course, one may often want to use a threshold parameter other than the ~~mean~~ median of the marginal ~~distribution~~distributions. The choice of τ is somewhat application specific; for example, it can be chosen such as to focus on high precipitation amounts. Thus, it is important that the FTE ~~metric~~ histogram maintains discrimination ability for different choices of τ. For a visual example, the same experiment depicted in Fig. 3 is repeated in Fig. 4, but with FTE histograms constructed using $\tau = 2$ (equivalent to two standard deviations from the mean in this case). When the ensemble fields have a correlation length that is slightly too small (row A), the resulting FTE histogram is ∪-shaped and has a slight right-skew due to the higher threshold, but correctly indicates excessive spatial variability. When the ensemble exhibits insufficient spatial variability, i.e. correlation length is slightly too large (row C), the FTE histogram is ∩-shaped and somewhat left-skewed. Reassuringly, the FTE histogram remains flat when the ensemble fields share the same correlation length as the verification fields (row B). While these results are in agreement with Table 1, the effect of the threshold can be studied more generally using the estimated ~~beta parameters~~ β-score and β-bias.

Figure 4. ~~Same as~~ As Fig. 3, but ~~for~~ with threshold $\tau = 2$.

3.3.2 Quantifying deviation from uniformity

Recall that we propose quantifying the shape of the FTE histogram ~~by the pair of beta distribution parameters (a, b) that approximate its shape. When $a = b = 1$~~with the β-score and β-bias. When $\beta_S = \beta_B = 0$, the FTE histogram is perfectly uniform. How do these ~~parameter values~~ metrics change as the threshold τ increases? Figure 5 ~~shows that when the correlation length~~ demonstrates how the β-score and β-bias vary over increasing thresholds for different correlation length ratios; the estimated β-distribution parameters are also depicted for comparison with Table 1. Where provided, confidence intervals were estimated via the the nonparametric bootstrap method (see Delignette-Muller and Dutang, 2015; Cullen and Frey, 1999). When the correlation length ratio is 1.0, ~~the estimated beta distribution parameters are both about 1~~ both the β-score and β-bias are approximately zero for every choice of τ, correctly indicating ~~flat histograms~~a spatially calibrated ensemble. When the correlation length ratio is less (greater) than 1.0, the ~~beta parameters~~ β-scores are themselves generally less ~~than 1, indicating $\cup$-shaped histograms. Conversely, when the ratio is greater than 1.0,~~ (greater) than zero, indicating excessive (insufficient) spatial variability. As previously discussed, the ~~beta parameters are themselves greater than 1, indicating $\cap$-shaped histograms.~~

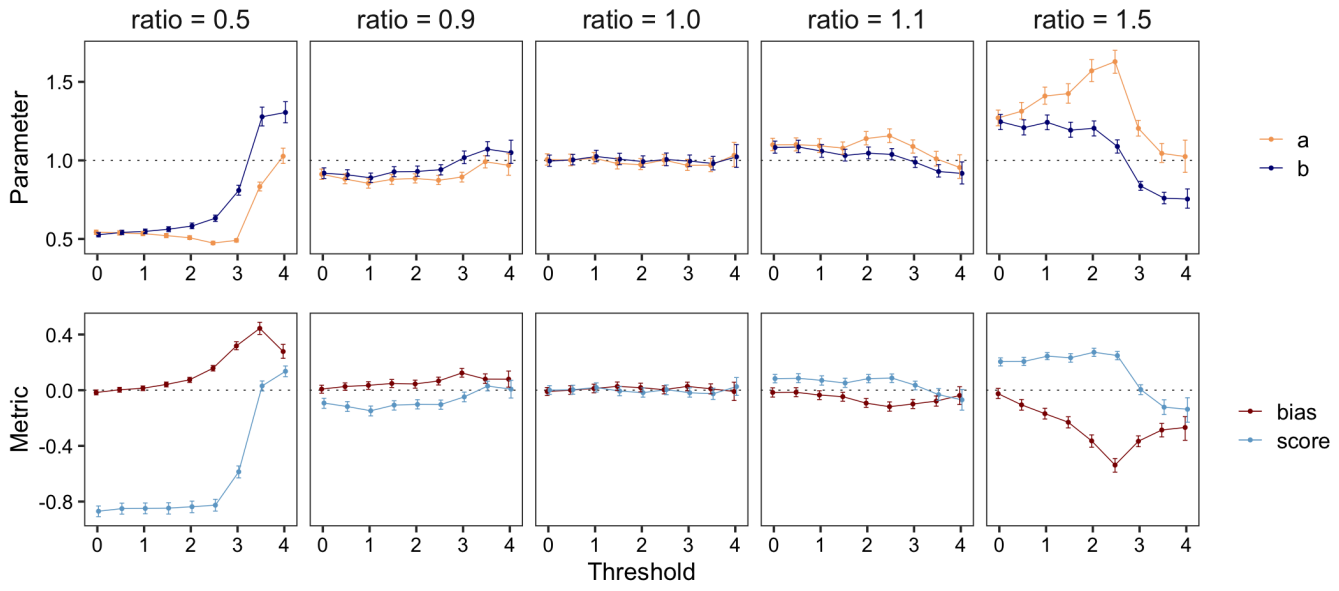

Figure 5. Estimated beta distribution parameters (top) and corresponding β-score and β-bias (bottom) of FTE histograms calculated over different thresholds for forecasts with low, even, and high correlation length ratios about $a_0 = 2$. Vertical lines denote the 95% confidence interval found via the nonparametric bootstrap method.

There is an interesting tendency toward unity in the beta parameters for high thresholds β-bias becomes more pronounced at higher thresholds, thus highlighting the inextricable link between threshold and skewness. Above a very high threshold of about three standard deviations (i.e., ~~τ greater than three standard deviations from the mean)~~$\tau \approx 3$), both metrics exhibit a tendency toward zero. This is ~~because the~~ partially due to the fact that the number of exceedances for high thresholds will often be zero ~~. Since~~ and since ranks are only discarded from the histogram if all ensemble members and the verification field have the same FTE, the FTE histogram for very high thresholds will be composed largely of ranks resulting from ties broken uniformly at random. This results in a more deceptively uniform histogram which ~~is why the estimated beta parameters tend toward 1 for high thresholds .~~ explains the tendency toward zero. For the most extreme thresholds studied here, the confounding effect of resolving ties (which can exist between all but a single ensemble member FTE) at random becomes very dominant, and histogram shapes get distorted to a degree where the interpretations provided in Table 1 no longer hold. The associated FTE histograms still exhibit non-uniformity and thus indicate that the ensemble forecasts are not perfectly calibrated, but it becomes impossible to diagnose the particular type of miscalibration from the histogram shape. We can also see that the sampling variability increases with increasing threshold since more and more uninformative cases with fully tied FTE values exist, so a much larger total number of verification cases is required in order to have a comparable number of informative cases. In practice, if FTE histograms are used as a diagnostic tool, we recommend focusing on moderate thresholds. If they are used to compare the calibration of different forecast systems, they can still be effective at more extreme thresholds.

Another variable of interest in evaluating the FTE ~~metric~~ histograms is the size of the domain to which the metric is applied. In our simulation framework, making the domain larger or smaller while keeping the correlation length constant is equivalent to keeping the domain size constant and varying the correlation length of the verification field. That is, for a fixed domain size, a smaller correlation length mimics a "large domain" (with low resolution) and larger correlation length mimics a "small domain~~." Analyzing estimated beta parameters~~" (with high resolution). Analyzing the β-score and β-bias over a range of correlation length ratios is then equivalent to studying the FTE ~~metric' s~~ histograms' utility for different domain sizes. For the domain used in this study, a correlation length of 1 is considered small and 3 is considered large ~~.~~ (see Fig. 2). In either case, Fig. 6 shows that the ~~beta parameters estimated using the FTE metric quickly rise above or fall below unity~~ β-score quickly deviates from zero when the correlation length ~~ratio is not 1.0 itself. The steep slopes~~ *ratio* is different from one. The β-score's relatively steep slope around the correlation length of 1.0 in both cases ~~indicate~~ indicates that the FTE ~~metric~~ histogram maintains good discrimination ability regardless of domain size~~.~~, provided that there are sufficiently many grid points within the domain to keep the (spatial) sampling variability associated with the calculation of the FTE values low. Notably, the inverse relationship between β-bias and the correlation length ratio is also in agreement with Table 1.

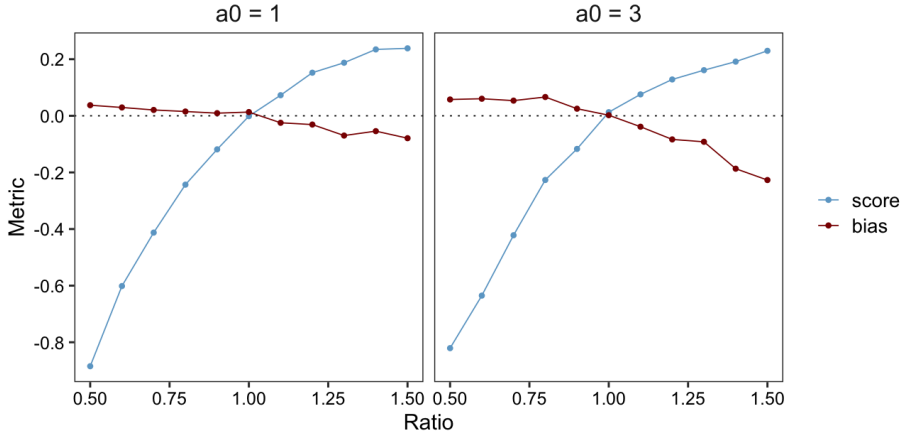

Figure 6. Estimated ~~beta parameters~~ β-score and β-bias of FTE histograms constructed with $\tau = 1$ for verification correlation length ~~$a_0 = 1, 3$~~ $a_0 \in \{1, 3\}$ and ensemble range a_M varying from $0.5a_0$ to $1.5a_0$~~, with $\tau = 1$~~.

We now turn attention back to our motivating figure (Fig. 1), which was created with verification correlation length $a_0 = 2$. Forecast 1 exhibited the correct spatial structure (i.e., $a_M = 2$), forecast 2 was incorrectly specified with correlation length a_M = 3, and forecast 3 was incorrectly specified with correlation length a_M = 1.8. While it may be obvious that forecast 2 has incorrect spatial structure, the structural difference between forecasts 1 and 3 is not so apparent. However, as demonstrated by the analysis above (Fig. 2 – Fig.~~4~~ 5), these misspecifications are certainly identifiable using ~~the FTE metric~~FTE histograms.

4 Application to downscaling of ensemble precipitation forecasts

Distributed hydrological models like NOAA's National Water Model (NWM) require meteorological inputs at a relatively high spatial resolution. At shorter forecast lead times (typically up to one or two days ahead) limited-area NWP models provide such high-resolution forecasts, but for longer lead times only forecasts from global ensemble forecast systems like NOAA's GEFS are available. These come at a relatively coarse resolution and need to be downscaled (statistically or dynamically) to the high-resolution output grid. Here, we use a combination of the statistical post-processing algorithm proposed by Scheuerer and Hamill (2015), ensemble copula coupling (ECC; Schefzik et al., 2013), and the spatial downscaling method proposed by Gagnon et al. (2012) to obtain calibrated, high-resolution precipitation ~~forecasts~~ forecast fields based on GEFS ensemble forecasts. Does the spatial disaggregation method produce precipitation fields with appropriate sub-grid scale variability? This question will be answered using the FTE-based verification metric discussed above.

4.1 Data and downscaling methodology

We consider 6-hour precipitation accumulations over a region in the South-Eastern US between -91° and -81° longitude and 30° and 40° latitude during the period from January 2002 to December 2016. Ensemble precipitation forecasts for lead time 66-h to 72-h were obtained from NOAA's second-generation GEFS reforecast dataset (Hamill et al., 2013) at a horizontal resolution of $\sim$0.5°. Downscaling and verification is performed against precipitation analyses from the $\sim$0.125° climatology-calibrated precipitation analysis (CCPA) dataset (Hou et al., 2014).

In order to obtain calibrated ensemble precipitation forecasts at the CCPA grid resolution we proceed in three steps. First, we apply the post-processing algorithm by Scheuerer and Hamill (2015) to the GEFS forecasts and upscaled (to the GEFS grid resolution) precipitation analyses in order to remove systematic biases and ensure adequate representation of forecast uncertainty at this coarse grid scale. The resulting predictive distributions are turned back into an 11-member ensemble using the ECC-mQ-SNP variation (Scheuerer and Hamill, 2018) of the ECC technique. This variation removes discontinuities and avoids randomization that can occur when the standard ECC approach is applied to precipitation fields. Finally, each ensemble member is downscaled from the GEFS to the CCPA grid resolution using a slightly simplified version of the Gibbs sampling disagregation model (GSDM) proposed by Gagnon et al. (2012). To generate downscaled fields with spatial properties that vary depending on the season, we rely here on a monthly calibration of the GSDM, rather than on meteorological predictors as in the original model. The 15 years of data are cross-validated: one year at a time is left out for verification and the post-processing and downscaling models are fitted with data from the remaining 14 years. Repeating this process for all years leaves us with 15 years of downscaled ensemble forecasts and verifying analyses. See Fig. 7 for a visual reference.

4.2 Univariate verification

Before applying the FTE histogram to investigate whether the spatial disaggregation used in the downscaling method produces precipitation fields with appropriate sub-grid scale variability, we check the calibration of the univariate ensemble forecasts across all fine scale grid points. We study (separately) the months January, April, July, and October in order to represent winter,

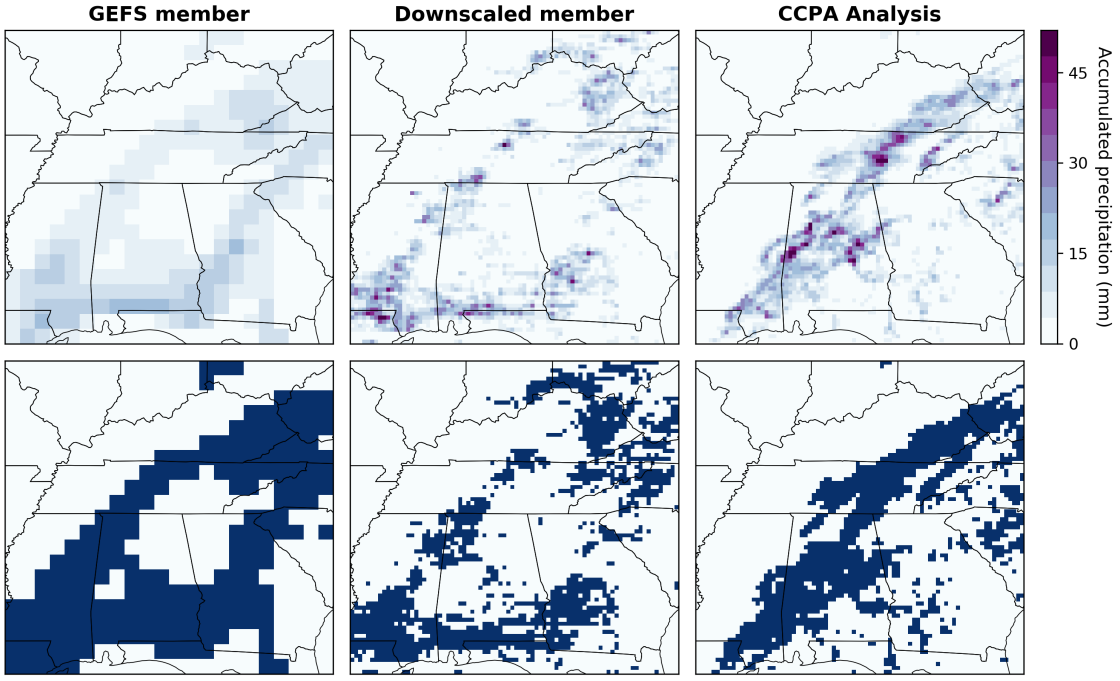

Figure 7. Examples of different data fields for 6-hour precipitation accumulation on July 24, ~~2004.~~ 2004 (top) and corresponding 5mm binary exceedance fields (bottom; dark blue regions indicate threshold exceedance). From left to right: coarse-scale GEFS ensemble member, the same member downscaled to the analysis resolution, and the corresponding CCPA analysis.

spring, summer, and fall, respectively. Daily analyses and corresponding ensemble forecasts from each of these months are pooled over the entire verification period and all grid points within the study area, and are used to construct the verification rank histograms in Fig. ~~8. Note that grid points receiving the same rank for the observation and all forecast fields~~ 8. Cases where all ensemble member forecasts and the analysis are tied – for example, when there is zero accumulation at a grid point for all fields – are withheld from the histogram to avoid artificial uniformity introduced by breaking ties in rank at random.

~~Recall that a calibrated forecast should yield~~ Ideally, the statistical post-processing and downscaling should yield calibrated ensemble forecasts and thus rank histograms that are approximately uniform. Clearly, the rank histograms for the downscaled forecast fields shown in Fig. 8 are not ~~exactly~~ uniform; there is a consistent peak in the higher ranks indicating that the downscaled ensemble forecasts tend to underestimate precipitation accumulations, especially in fall and winter. This bias could be an indication that either the post-processing distribution (gamma) or the disaggregation distribution (log-normal) are not perfectly suited to represent the respective forecast uncertainties. It may also be a result of a superposition of biases in different sub-domains or for different weather situations. Univariate calibration in July - which happens to be a month with more frequent precipitation in this region of the US - is relatively good, and while the histograms of other months show clear departures from uniformity, there is at least no strong ∪– or ∩–shape to indicate significant dispersion errors. We thus continue

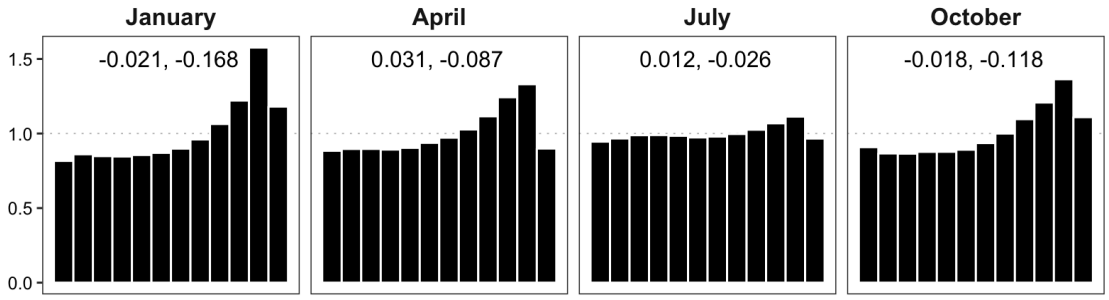

Figure 8. Verification rank histograms (density) for downscaled fields at representative months with cases of fully tied ranks removed. Estimated β-score (left) and β-bias (right) annotated.

with our analysis of the spatial calibration of the downscaled ensemble ~~forecasts~~ forecast fields, keeping in mind though that ~~some of the possible non-uniformity of FTE histogramsin January, April and October could be due to univariate miscalibration.~~ the under-forcast biases seen in Fig. 8 will carry over to the FTE histograms, and will superimpose any shape resulting from spatial miscalibration. In July, where only a mild under-forcast bias is observed, we will have the best chance of drawing unambiguous conclusions about the spatial structure from the shape of the FTE histogram.

4.3 Verification of spatial structure

In the remaining analysis, we employ FTE histograms to investigate the spatial properties of the ensemble forecast fields obtained by the downscaling algorithm for the same representative months outlined above. Spatial variability of precipitation fields depends on whether precipitation is stratiform or convective, and in the latter case also on the type of convection (local vs synoptically forced). The frequency of occurrence of these categories has a seasonal cycle, and it is therefore interesting to study how well the downscaling methodology works in different seasons. The first step in computing the FTE is deciding what value to use for the threshold. If the climatology varies strongly across the domain, it may be desirable to use a variable threshold such as a climatology percentile. However, the South-Eastern US is a flat and relatively homogeneous region meaning the precipitation accumulation patterns will not be affected as much by orography, and we therefore select a fixed threshold for constructing ~~the FTE~~ FTE histograms. Another advantage of this approach is that a fixed threshold has a direct physical interpretation; here we use thresholds of 5mm, 10mm, and 20mm to study the spatial calibration of the ensemble for low, medium, and high accumulation levels over the 6-hour window.

In Fig. 9, it is clear by visual inspection that the FTE histograms are all $\cup$-shaped to some extent, though the ~~fitted beta parameters~~ corresponding β-scores highlight that the histograms are explicitly more uniform in the fall and winter months. In the spring and summer months (i.e., April and and July) the histograms reveal a clear under-dispersion in the ensemble FTEs at all analyzed thresholds. ~~Since we noted above that July (and, to a lesser extent, April) had the best univariate calibration, this~~ This would suggest that the downscaled ensemble overestimates fine scale variability during the seasons with more convective

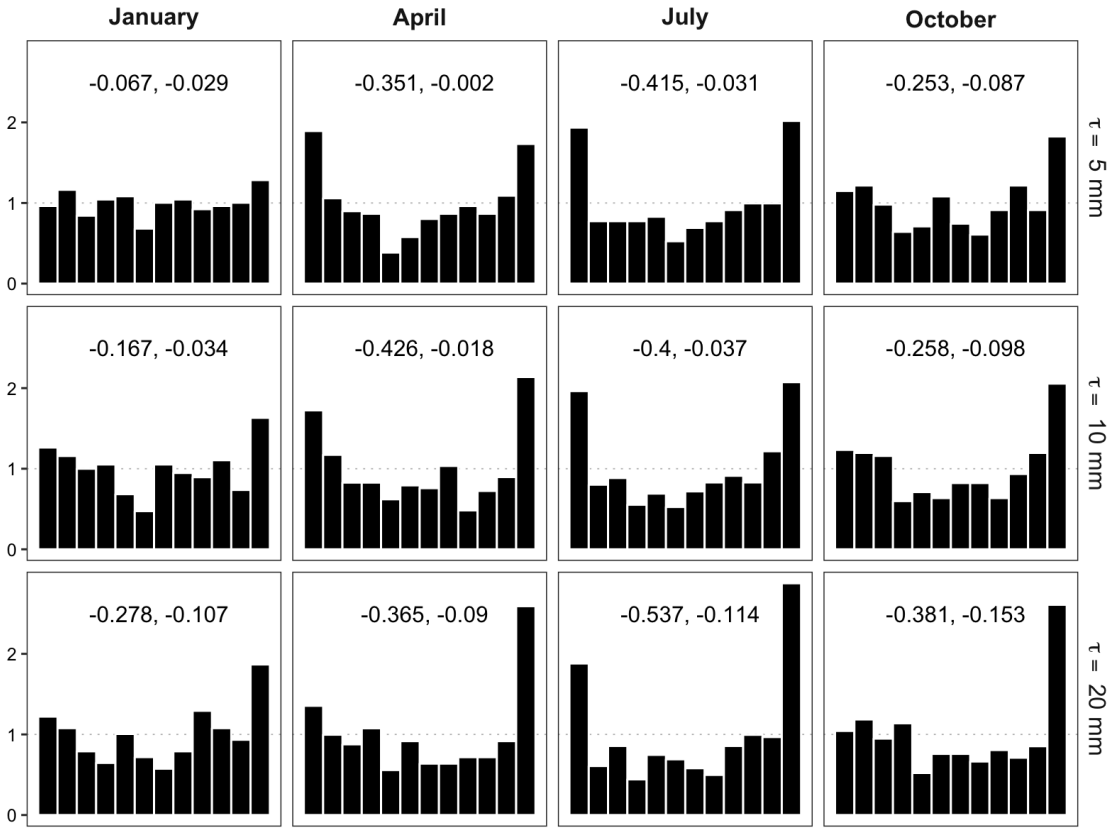

Figure 9. FTE histograms for downscaled fields at different thresholds in representative months~~with estimated beta distribution parameters~~. Estimated β-score (left) and β-bias (right) annotated.

events. This could indicate that the calibration procedure of the GSDM downscaling method in Gagnon et al. (2012) struggles with selecting good parameters that produce downscaled precipitation fields with just the right amount of spatial variability during the summer season with mainly (but not exclusively) convective precipitation. The FTE ~~metric~~ histograms can thus provide valuable diagnostic information that helps identify shortcomings of a forecast methodology. Indeed, in one of our current projects we seek to improve the GSDM, with one objective being to calibrate the model such that the downscaled fields reproduce the correct amount of spatial variability, in a flow-dependent fashion, using meteorological predictors such as instability indices and vertical wind shear (Bellier et al., 2020). For the β-biases seen in Fig. 9, the interpretation is more difficult. Their value at higher thresholds has the opposite sign as what we would expect from Table 1 in a situation where the forecast fields simulate excessive fine scale variability. As noted above, however, this is likely due to an under-forecast bias in the marginal distributions and the associated effect on the β-bias which is opposite to (and seems to be dominating) the effects of spatial miscalibration.

5 Conclusions

When forecasting meteorological variables on a spatial domain, it is important for many applications that not only the marginal forecast distributions but also the spatial (and/or temporal) correlation structure is represented adequately. In some instances, misrepresentation of spatial structure by ensemble forecast fields may be visually obvious; otherwise, a quantitative verification metric is desired to objectively evaluate the ensemble calibration. The FTE metric studied here is a projection of a multivariate quantity (i.e., a spatial field) to a univariate quantity, and can be combined with the concept of a (univariate) verification rank histogram to analyze the spatial structure of ensemble forecast fields. This idea has first been applied by Scheuerer and Hamill (2018) to study the properties of downscaled ensemble precipitation forecasts, but an understanding of the general capability of the FTE metric to detect misrepresentation of the spatial structure by the ensemble has been lacking as yet.

In this paper, we performed a systematic study in which we simulated ensemble forecast and verification fields with different correlation lengths to understand how well a misspecification of the correlation length can be detected by the FTE metric. To this end, the metric was slightly extended and is composed of three steps: (1) calculate the FTE of each verification and ensemble forecast field, (2) construct an FTE histogram over available instances of forecast and verification times, and (3) ~~fit beta parameters to the~~ derive the β-score and β-bias from the stochastically disaggregated FTE histogram ~~by maximum likelihood to summarize~~ to characterize departure from uniformity. We have found that the FTE metric is capable of detecting even minor issues with the correlation length (e.g., 10% miscalibration) in ensemble forecasts, and this conclusion was consistent across a range of thresholds and domain sizes. Applied in a data example with downscaled precipitation forecast fields, the FTE metric pointed to some shortcomings of the underlying spatial disaggregation algorithm during the seasons where precipitation is ~~to be~~ driven by local convection.

The FTE metric is relatively simple and ~~thus~~ enjoys an easy and intuitive interpretation. In particular, the ~~estimated beta distribution parameters~~ β-score and β-bias can be compared according to Table 1 ~~to obtain a simple objective summary of dispersion and bias in the ensemble forecast.~~ to diagnose shortcomings in the calibration of ensemble forecasts. If different types of miscalibration occur together, additional diagnostic tools like univariate verification rank histograms have to be considered alongside the FTE histograms to disentangle the different effects. While we have focused on ~~verification rank~~ histograms in the analysis of the ~~univariate~~ verification FTE rank here, the same projection could also be used in combination with proper scoring rules. We believe that FTE histograms are a useful addition to the set of spatial verification metrics. They complement metrics like the wavelet-based verification approach proposed by Buschow et al. (2019) which has additional capabilities when it comes to analyzing aspects of the spatial texture of forecast fields but is not primarily targeted at proper uncertainty quantification by an ensemble.

Code and data availability. The code and data used for this study are available in the accompanying Zenodo repositories (code: https://doi.org/10.5281/zenodo.3592506, data: https://doi.org/10.5281/zenodo.3592495). The most recent version of the code can also be found in JJ's Git repository (https://github.com/joshhjacobson/FTE).

Appendix A: Stochastic disaggregation of transformed ranks

The data vector of ranks ~~R~~ r has discrete elements $r_i \in \{1, 2, \ldots, k+1\}$ where k is the number of ensemble members. In order to disaggregate these elements to a continuous domain for use with maximum likelihood estimation, the following algorithm is applied to each element r_i:

1. Let $d_i = (r_i - \frac{1}{2})/(k+1)$.
2. Simulate a (continuous) uniform random variable

$$U_i \sim Uniform\left(d_i - \frac{1}{2(k+1)}, d_i + \frac{1}{2(k+1)}\right)$$

3. Set $r_i = U_i$.

The effect of Step 1 is a mapping into $[0,1]$, while Step 2 is the stochastic disaggregation to evenly-spaced uniform intervals whose supports form a partition of unity of [0,1].

Appendix B: ~~Ensemble~~ Properties of simulated ensemble members~~with standard Gaussian marginals~~

~~Suppose X and W~~

Let $Z_M(s)$ and $W_i(s)$ be independent, mean-zero Gaussian processes, each with Matérn covariance function $M(d|\nu_M, a_M)$. Now suppose Z_M and W_i are independent standard Gaussian random variables ~~. Letting~~ representing the marginal distribution of processes $Z_M(s)$ and $W_i(s)$. Setting random variable

$$Z_i = \omega X Z_M + \sqrt{1-\omega^2} W_i, \quad \omega \in [-1, 1], \tag{B1}$$

we see ~~$\mathbb{E}[Z] = 0$~~ $\mathbb{E}[Z_i] = 0$ by linearity of the expectation operator, and

$$\begin{aligned} \mathrm{Var}[Z]\mathrm{Var}[Z_i] &= \mathrm{Var}[\omega X + \sqrt{1-\omega^2} W]\mathrm{Var}[\omega Z_M + \sqrt{1-\omega^2} W_i] \\ &= \omega^2 \mathrm{Var}[X]\mathrm{Var}[Z_M] + (1-\omega^2)\mathrm{Var}[W]\mathrm{Var}[W_i] \\ &= 1 \end{aligned} \tag{B2}$$

using independence of ~~X and W.~~ Z_M and W_i. Then Z_i is a standard Gaussian random variable representing the marginal distribution of ensemble member $Z_i(s) = \omega Z_M(s) + \sqrt{1-\omega^2} W_i(s)$. Further, observe that

$$\begin{aligned}
&\text{Cov}[Z_i(s+h), Z_i(s)] \\
&\quad = \text{Cov}[\omega Z_M(s+h) + \sqrt{1-\omega^2} W_i(s+h), \omega Z_M(s) + \sqrt{1-\omega^2} W_i(s)] \\
&\quad = \omega^2 \text{Cov}[Z_M(s+h), Z_M(s)] + (1-\omega^2)\text{Cov}[W_i(s+h), W_i(s)] \\
&\quad = \omega^2 M(||h||, \nu_M, a_M) + (1-\omega^2) M(||h||, \nu_M, a_M) \\
&\quad = M(||h||, \nu_M, a_M)
\end{aligned} \tag{B3}$$

by independence of $Z_M(s)$ and $W_i(s)$. That is, ensemble members $Z_i(s)$, $i = 1, \ldots, k$, preserve the covariance structure of the ensemble mean $Z_M(s)$.

Appendix C: Derivation of an appropriate co-located correlation coefficient

Suppose Z_0 and Z_M are standard Gaussian random variables with $\text{Corr}[Z_0, Z_M] = \rho$, and $\{W_i\}_{i=1}^k$ is a set of independent standard Gaussian random variables, each independent of Z_0 and Z_M. Define

$$Z_i = \omega Z_M + \sqrt{1-\omega^2} W_i, \quad i = 1, \ldots, k, \quad \omega \in [-1, 1]. \tag{C1}$$

From Appendix B we have that each of Z_i is again a standard Gaussian random variable. Then, for $i \neq j$, we see that

$$\begin{aligned}
\text{Cov}[Z_i, Z_j] &= \text{Cov}(\omega Z_M + \sqrt{1-\omega^2} W_i, \omega Z_M + \sqrt{1-\omega^2} W_j) \\
&= \omega^2 \text{Cov}(Z_M, Z_M) \\
&= \omega^2
\end{aligned} \tag{C2}$$

using pairwise independence of Z_M, W_i and W_j. Using a similar technique we see

$$\text{Cov}[Z_0, Z_i] = \omega\rho. \tag{C3}$$

Now let $\mathbf{Z} = (Z_0, Z_1, \ldots, Z_k)'$. Then,

$$\text{Cov}[\mathbf{Z}] = \begin{pmatrix} 1 & \omega\rho & \ldots & \ldots & \omega\rho \\ \omega\rho & \ddots & \omega^2 & \ldots & \omega^2 \\ \vdots & \omega^2 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \omega^2 \\ \omega\rho & \omega^2 & \ldots & \omega^2 & 1 \end{pmatrix} \tag{C4}$$

Setting $\rho = \omega$ is thus necessary (except for the trivial case where $\omega = 0$) and sufficient for univariate probabilistic calibration of the ensemble as this choice makes Z_0 indistinguishable from $Z_1, \ldots, Z_k$ in distribution.

Appendix D: Sensitivity of FTE histograms to forecast skill

The simulation setup introduced in section 3.2 allows us to control the skill of the synthetic ensemble forecasts through the parameters ρ and ω, where (as shown above) the restriction $\rho = \omega$ is required to ensure calibration of the marginal distributions. Does the sensitivity of the FTE histogram to mis-specified correlation lengths change with changing forecast skill? To investigate this further, we show results for the extreme 'no skill' case where $\rho = \omega = 0$, to complement those shown above where we simulated forecasts with a relatively high correlation ($\rho = \omega = 0.8$) between ensemble mean and verification at each grid point. The other parameters remain unchanged; i.e., simulation experiments are performed for $a_0 = 2$ and $a_M \in \{1, 1.8, 2, 2.2, 3\}$. By Eq. (8), the ensemble members are then independent realizations of a mean zero Gaussian random field with Matérn covariance having correlation length $a = a_M$.

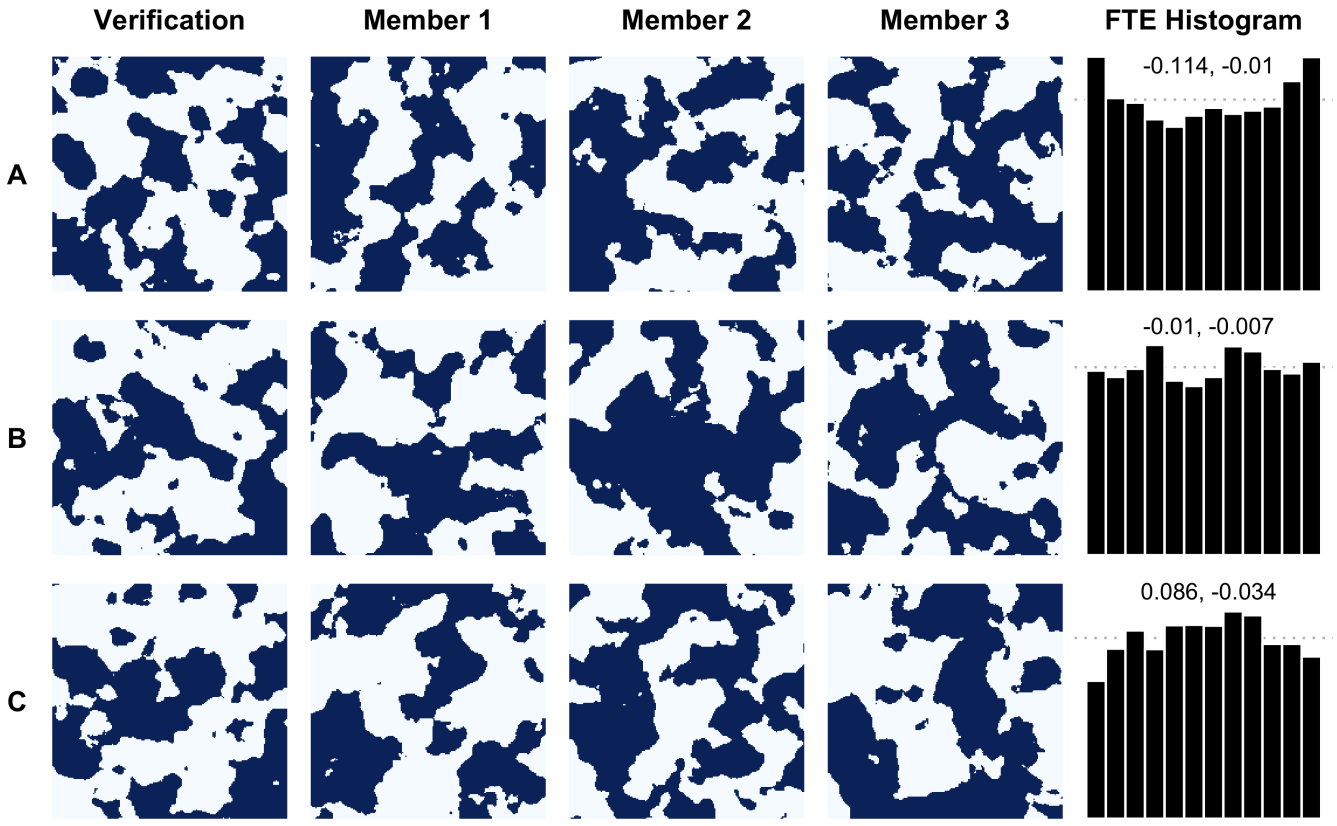

Figure D1. As Fig. 3, but ensemble fields are constructed with skill parameter $\rho = \omega = 0$.

Figure D1 gives an example of simulated ensemble member fields which are mutually independent and uncorrelated with the verification field, while their spatial structure is 10% miscalibrated in the case of rows A and C. In contrast to Fig. 3, where the positive skill of the ensemble forecasts entails some degree of correspondence between the features in the forecast and

verification fields, there is no such correspondence in between the fields in Figure D1. The corresponding FTE histograms and their associated β-scores, however, are able to identify row B as spatially calibrated and row A (row C) as having a correlation length ratio that is too small (large).

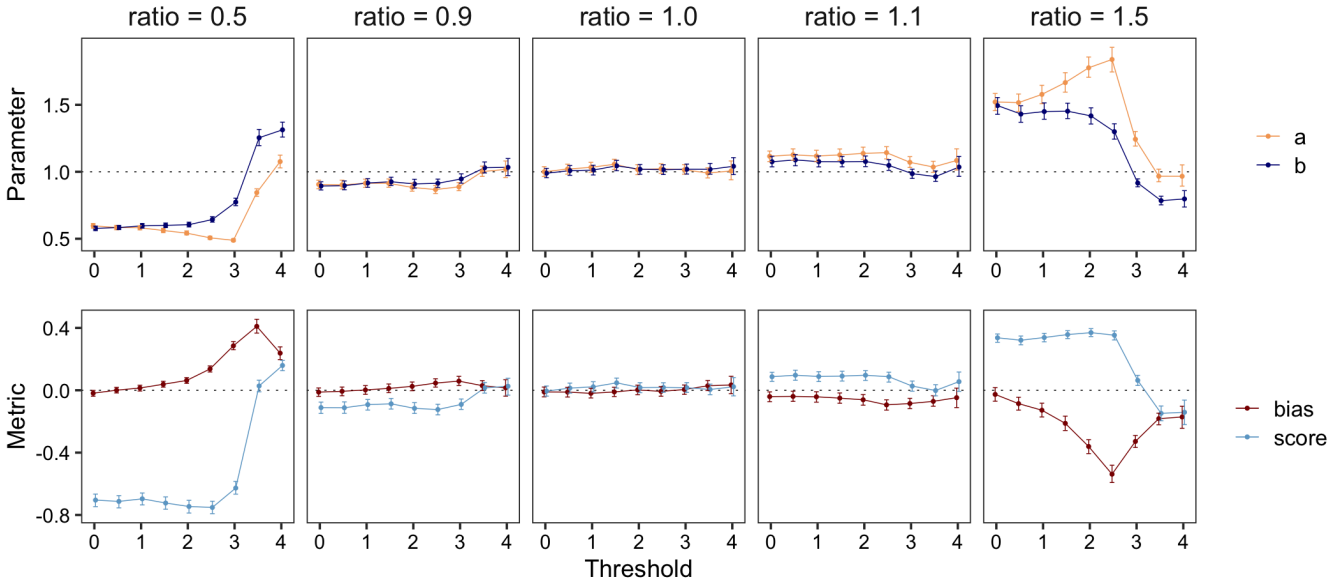

Figure D2. As Fig. 5, but ensemble fields are constructed with skill parameter $\rho = \omega = 0$.

A similar story is provided by Fig. D2 which, like Fig. 5 where $\rho = \omega = 0.8$, demonstrates how the estimated β-distribution parameters, β-score, and β-bias vary over increasing thresholds for different correlation length ratios. While the associated experiments differ in that ensemble members have no univariate skill here (i.e., $\rho = \omega = 0$), the behavior witnessed in the two figures is nearly indistinguishable. Thus, we conclude that correlation between the ensemble and verification has a negligible effect on the FTE histograms' ability to detect miscalibration in spatial structure. This was not obvious to us a priori, but perhaps one can think of these 'no skill' simulations as the residual fields that remain after the 'predictable component' has been subtracted from both ensemble member and verification fields. In any case, the insensitivity to forecast skill is good news for the practical application of FTE histograms, where forecast skill is usually unknown, and confounding effects are undesirable. It is a reminder though that they are a tool for assessing forecast calibration, not forecast skill.

Author contributions. This study is based on the Master's work of JJ, under supervision of WK and MS. The concept of this study was developed by MS and extended upon by all involved. JJ implemented the study and performed the analysis with guidance from WK and MS. JB provided the downscaled GEFS forecast fields. JJ, WK, and MS collaborated in discussing the results and composing the manuscript, with input from JB on section 4.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. JJ was supported by NSF DMS-1407340. WK was supported by NSF DMS-1811294 and DMS-1923062. MS was supported by funding from NOAA/ESRL/PSD. JB was supported by a funding agreement between NOAA/ESRL/PSD and NOAA/NWS/MDL on the development and transfer of advanced statistically postprocessed probabilistic ensemble guidance. This work utilized resources from the University of Colorado Boulder Research Computing Group, which is supported by the NSF (awards ACI-1532235 and ACI-1532236), the University of Colorado Boulder, and Colorado State University.

References

Anderson, J. L.: A method for producing and evaluating probabilistic forecasts from ensemble model integrations, Journal of Climate, 9, 1518–1530, https://doi.org/10.1175/1520-0442(1996)009<1518:AMFPAE>2.0.CO;2, 1996.

Apanasovich, T. V., Genton, M. G., and Sun, Y.: A valid Matérn class of cross-covariance functions for multivariate random fields with any number of components, Journal of the American Statistical Association, 107, 180–193, https://doi.org/10.1080/01621459.2011.643197, 2012.

Bellier, J., Scheuerer, M., and Hamill, T. M.: Precipitation downscaling with Gibbs sampling: An improved method for producing realistic, weather-dependent and anisotropic fields, Journal of Hydrometeorology, under review, 2020.

Buizza, R., Bidlot, J.-R., Wedi, N., Fuentes, M., Hamrud, M., Holt, G., and Vitart, F.: The new ECMWF VAREPS (variable resolution ensemble prediction system), Quarterly Journal of the Royal Meteorological Society, 133, 681–695, https://doi.org/10.1002/qj.75, 2007.

Buschow, S. and Friederichs, P.: Using wavelets to verify the scale structure of precipitation forecasts, Adv. Stat. Clim. Meteorol. Oceanogr., 6, 13–30, https://doi.org/10.5194/ascmo-6-13-2020, 2020.

Buschow, S., Pidstrigach, J., and Friederichs, P.: Assessment of wavelet-based spatial verification by means of a stochastic precipitation model (wv_verif v0.1.0), Geoscientific Model Development, 12, 3401–3418, https://doi.org/10.5194/gmd-12-3401-2019, 2019.

Cullen, A. C. and Frey, H. C.: Probabilistic techniques in exposure assessment, Plenum Press, USA, 1999.

Delignette-Muller, M. L. and Dutang, C.: fitdistrplus: An R package for fitting distributions, Journal of Statistical Software, 64, 1–34, https://doi.org/10.18637/jss.v064.i04, 2015.

Feldmann, K., Scheuerer, M., and Thorarinsdottir, T. L.: Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression, Monthly Weather Review, 143, 955–971, https://doi.org/10.1175/MWR-D-14-00210.1, 2015.

Gagnon, P., Rousseau, A. N., Mailhot, A., and Caya, D.: Spatial disaggregation of mean areal rainfall using Gibbs sampling, Journal of Hydrometeorology, 13, 324–337, https://doi.org/10.1175/JHM-D-11-034.1, 2012.

Genton, M. G. and Kleiber, W.: Cross-covariance functions for multivariate geostatistics, Statistical Science, 30, 147–163, https://doi.org/10.1214/14-STS487, 2015.

Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., and Ebert, E. E.: Intercomparison of spatial forecast verification methods, Weather and Forecasting, 24, 1416–1430, https://doi.org/10.1175/2009WAF2222269.1, 2009.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., and Johnson, N. A.: Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds, TEST, 17, 211, https://doi.org/10.1007/s11749-008-0114-x, 2008.

Gneiting, T., Kleiber, W., and Schlather, M.: Matérn cross-covariance functions for multivariate random fields, Journal of the American Statistical Association, 105, 1167–1177, https://doi.org/10.1198/jasa.2010.tm09420, 2010.

Hamill, T. M.: Interpretation of rank histograms for verifying ensemble forecasts, Monthly Weather Review, 129, 550–560, https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2, 2001.

Hamill, T. M., Bates, G. T., Whitaker, J. S., Murray, D. R., Fiorino, M., Galarneau, T. J., Zhu, Y., and Lapenta, W.: NOAA's second-generation global medium-range ensemble reforecast dataset, Bulletin of the American Meteorological Society, 94, 1553–1565, https://doi.org/10.1175/BAMS-D-12-00014.1, 2013.

Hou, D., Charles, M., Luo, Y., Toth, Z., Zhu, Y., Krzysztofowicz, R., Lin, Y., Xie, P., Seo, D.-J., Pena, M., and Cui, B.: Climatology-calibrated precipitation analysis at fine scales: Statistical adjustment of stage IV toward CPC gauge-based analysis, Journal of Hydrometeorology, 15, 2542–2557, https://doi.org/10.1175/JHM-D-11-0140.1, 2014.

Kapp, F., Friederichs, P., Brune, S., and Weniger, M.: Spatial verification of high-resolution ensemble precipitation forecasts using local wavelet spectra, Meteorologische Zeitschrift, 27, 467–480, https://doi.org/10.1127/metz/2018/0903, 2018.

Keller, J. D. and Hense, A.: A new non-Gaussian evaluation method for ensemble forecasts based on analysis rank histograms, Meteorologische Zeitschrift, 20, 107–117, https://doi.org/10.1127/0941-2948/2011/0217, 2011.

Kleiber, W.: Coherence for multivariate random fields, Statistica Sinica, 27, 1675–1697, https://doi.org/10.5705/ss.202015.0309, 2017.

Leutbecher, M. and Palmer, T.: Ensemble forecasting, Journal of Computational Physics, 227, 3515–3539, https://doi.org/10.1016/j.jcp.2007.02.014, 2008.

Murphy, A. H. and Winkler, R. L.: Reliability of subjective probability forecasts of precipitation and temperature, Journal of the Royal Statistical Society. Series C (Applied Statistics), 26, 41–47, https://doi.org/10.2307/2346866, 1977.

Roberts, N. M. and Lean, H. W.: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events, Monthly Weather Review, 136, 78–97, https://doi.org/10.1175/2007MWR2123.1, 2008.

Schefzik, R., Thorarinsdottir, T. L., and Gneiting, T.: Uncertainty quantification in complex simulation models using ensemble copula coupling, Statistical Science, 28, 616–640, https://doi.org/10.1214/13-STS443, 2013.

Scheuerer, M. and Hamill, T. M.: Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted Gamma distributions, Monthly Weather Review, 143, 4578–4596, https://doi.org/10.1175/MWR-D-15-0061.1, 2015.

Scheuerer, M. and Hamill, T. M.: Generating calibrated ensembles of physically realistic, high-resolution precipitation forecast fields based on GEFS model output, Journal of Hydrometeorology, 19, 1651–1670, https://doi.org/10.1175/JHM-D-18-0067.1, 2018.

Schlather, M., Malinowski, A., Menck, P., Oesting, M., and Strokorb, K.: Analysis, simulation and prediction of multivariate random fields with package RandomFields, Journal of Statistical Software, Articles, 63, 1–25, https://doi.org/10.18637/jss.v063.i08, 2015.

Smith, L. A. and Hansen, J. A.: Extending the limits of ensemble forecast verification with the minimum spanning tree, Monthly Weather Review, 132, 1522–1528, https://doi.org/10.1175/1520-0493(2004)132<1522:ETLOEF>2.0.CO;2, 2004.

Thorarinsdottir, T. L., Scheuerer, M., and Heinz, C.: Assessing the calibration of high-dimensional ensemble forecasts using rank histograms, Journal of Computational and Graphical Statistics, 25, 105–122, https://doi.org/10.1080/10618600.2014.977447, 2016.

Toth, Z. and Kalnay, E.: Ensemble forecasting at NMC: The generation of perturbations, Bulletin of the American Meteorological Society, 74, 2317–2330, https://doi.org/10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2, 1993.

Wilks, D. S.: The minimum spanning tree histogram as a verification tool for multidimensional ensemble forecasts, Monthly Weather Review, 132, 1329–1340, https://doi.org/10.1175/1520-0493(2004)132<1329:TMSTHA>2.0.CO;2, 2004.

Zhou, X., Zhu, Y., Hou, D., Luo, Y., Peng, J., and Wobus, R.: Performance of the new NCEP Global Ensemble Forecast System in a parallel experiment, Weather and Forecasting, 32, 1989–2004, https://doi.org/10.1175/WAF-D-17-0023.1, 2017.

Ziegel, J. F. and Gneiting, T.: Copula calibration, Electronic Journal of Statistics, 8, 2619–2638, https://doi.org/10.1214/14-EJS964, 2014.