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

by

Joshuah Jacobson

B.S., University of Colorado Boulder, 2019

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Master of Science

Department of Applied Mathematics

2020

This thesis entitled:
Beyond Univariate Calibration: Verifying Spatial Structure in Ensembles of Forecast Fields
written by Joshuah Jacobson
has been approved for the Department of Applied Mathematics

_____

Prof. William Kleiber

_____

Dr. Michael Scheuerer

_____

Dr. Stephan Sain

Date _____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Jacobson, Joshuah (M.S., Applied Mathematics)

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

Thesis directed by Prof. William Kleiber

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. This work studies the properties of a new diagnostic tool designed for spatially-indexed ensemble forecast fields. The metric is based on a level-crossing statistic termed the fraction of threshold exceedance (FTE), and is calculated for the verification field, and separately for each ensemble member. The FTE 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 FTE histograms can 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.

Dedication

To my parents, Marty and Steffanie Jacobson. I am grateful for your support and inspired by your selfless spirit.

Acknowledgements

I am deeply indebted to a number of individuals and communities, without whom this endeavor would not have been possible. First, I would like to express particular appreciation for Prof. William Kleiber and Dr. Michael Scheuerer. In addition to their role as my advisors, they have been trusting and patient teachers whose contagious enthusiasm for research has afforded me great perspective. Thanks also to Joseph Bellier for providing the downscaled forecast fields as well as many valuable comments on this work. I am also grateful for Prof. Joseph Kazprzyk and Billy Raseman; their guidance and support through the Discovery Learning Apprenticeship provided me with an unbelievably smooth introduction to research. I would like to thank Anne Dougherty for her consistent encouragement and expertise as my academic advisor, for her tireless efforts in fostering the applied math community, and for securing funding for this work through the NSF EXTREEMS-QED grant DMS-1407340. The CU Boulder Research Computing group also has my thanks, their technical support and personable staff have been an invaluable resource. Special thanks is due to Scot Douglass and Mary Rader who have consistently challenged me to seek out the small details within the bigger picture. Finally, I would like to thank Greg Benton whose path has served as an inspiration for my own, as well as Amy DeCastro for her genuine compassion and many illuminating conversations.

Contents

Tables

Table

Figures

Figure

Chapter 1

Introduction

Ensemble prediction systems like the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble (Buizza et al., 2007) or the National Oceanic and Atmospheric Administration (NOAA) 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. 1.1: 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.

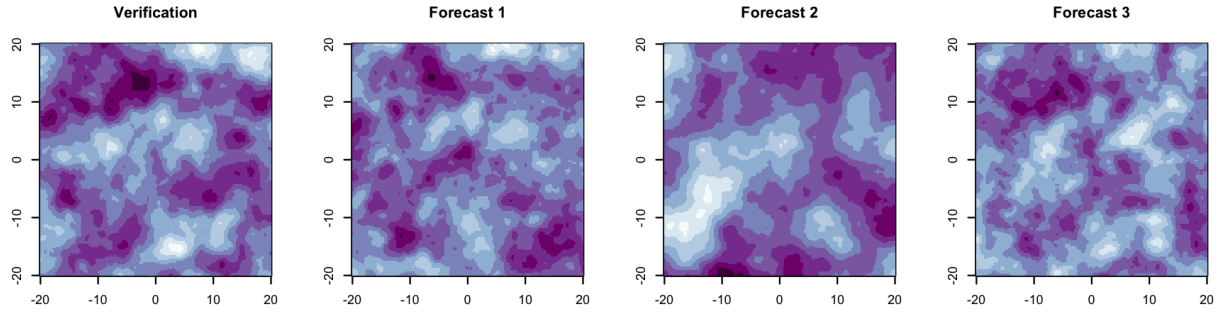

Figure 1.1: Simulated verification field and three associated forecast fields (arbitrary color scale) in which the spatial correlation length is either the same as for the verification, 10% miscalibrated, or 50% miscalibrated. Can you tell which is correct? Answer at the end of Chapter 3.

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 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. Our goal is similar but the approach studied here is more in line with 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 on the probabilistic nature of the forecasts. That is, we are studying whether the forecasts adequately represent spatial variability *as an ensemble*.

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) 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 work is to investigate this discrimination ability in detail.

In chapter 2, we describe the calculation of the FTE and the construction of the FTE histogram in depth. In chapter 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 chapter 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 chapter 5.

Chapter 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

$$\begin{aligned} \mathrm{FTE}(Z,\tau) &= \frac{1}{|D|}\int_D \mathbf{1}_{\{Z(s)>\tau\}}(s)ds \\ &= \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{\{Z(s)>\tau\}}(s_j) \end{aligned} \tag{2.1}$$

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. when the precipitation amount reported by the verification and predicted by all ensemble members is below the threshold τ everywhere in D.

Gathering ranks over N instances of forecast/verification pairs, $r_1, \ldots, r_N$, a natural way to communicate the rank behavior is through a histogram that we term the *FTE histogram* over the $k+1$ possible ranks. In construction, the FTE histogram is similar to the (univariate) rank histogram discussed in Anderson (1996) and Hamill (2001). That rank histogram, however, 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. 1.1). 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 $\cap$–shaped histogram can be expected; that is, 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, a $\cup$–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 (i.e., all forecast and verification FTEs are the same) are completely uninformative and can be discarded in calculating the FTE histogram in order to avoid artificial uniformity.

While the FTE histogram is visually intuitive, a quantitative measure for studying departures from uniformity is desirable. Akin to Keller and Hense (2011), we summarize the FTE histogram,

transformed to the unit interval, with two parameters from a beta distribution. However, as 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.1 for details. We then fit a beta distribution to the disaggregated ranks by maximum likelihood. This provides 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. In the ideal case, a and b are both exactly one, indicating that the FTE histogram is perfectly uniform. In practice, these parameters are never exactly one. The resulting set of possible parameter combinations and corresponding departures from uniformity are outlined in Table 2.1. With parameters a and b, we have obtained an easily interpreted measure of spatial forecast calibration.

Parameter Relationship	Histogram Interpretation
$a = b = 1$	Uniform
$a, b < 1$	$\cup$–shaped
$a, b > 1$	$\cap$–shaped
$a < b$	Right-skewed
$a > b$	Left-skewed

Table 2.1: Possible departures from histogram uniformity, as characterized by beta parameter relationships.

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 stochastically disaggregated FTE histogram by maximum likelihood to characterize departure from uniformity.

Chapter 3

Simulation study

In the present chapter we consider an extensive simulation study to assess the ability of the proposed FTE histogram in diagnosing mismatches between the correlation length of the forecast fields and that of the verification fields. It is not straightforward to define what a "correlation length" is in general for various meteorological quantities of interest such as precipitation and wind speeds, especially given possible heterogeneity and spatial nonstationarity over the study domain. However, the effect of using a threshold exceedance helps mitigate this problem. 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 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. 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

$$\begin{aligned} P(Z(s_i) > q(s_i)) &= P(F(Z(s_i)) > F(q(s_i))) \\ &= P(\Phi^{-1}(F(Z(s_i))) > \Phi^{-1}(F(q(s_i))) \\ &= P(X > \Phi^{-1}(0.9)) \end{aligned} \tag{3.1}$$

where X is a standard normal. Thus, we have shown that a field of random variables with possibly distinct local probability distributions can be transformed to Gaussian, and if we use local quantiles as the threshold then this is equivalent to a spatially-constant threshold on the transformed variables. 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

$$C_{ij}(h) = \mathrm{Cov}(Z_i(s+h), Z_j(s)) \tag{3.2}$$

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, i.e.

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=0}^{p}\sum_{\ell=0}^{p} a_i a_j C_{k\ell}(s_i - s_j) \geq 0 \tag{3.3}$$

for any choices of $a_i \in \mathbb{R}$ and $s_i \in D$, which simply means that any linear combination of $Z_0, \ldots, Z_k$ must have nonnegative variance (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

$$M(h|\nu, a) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{h}{a}\right)^{\nu} K_\nu \left(\frac{h}{a}\right) \tag{3.4}$$

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

$$C_{ii}(h) = \sigma_i^2 M(h|\nu_i, a_i), \qquad \text{for } i = 0, \ldots, k, \tag{3.5}$$

and

$$C_{ij}(h) = C_{ji}(h) = \rho_{ij}\sigma_i\sigma_j M(h|\nu_{ij}, a_{ij}), \qquad \text{for } 0 \leq i \neq j \leq k. \tag{3.6}$$

In this latter equation, $\rho_{ij} \in [-1, 1]$ 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 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, $k = 11$ (see Fig. 3.1 for an example).

The simulation setup follows a series of steps:

(1) Generate Z_0 and Z_M, the verification and ensemble mean as a mean zero bivariate Gaussian random field with multivariate Matérn spatial range 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 $a = a_M$ and smoothness $\nu = \nu_M = 1.5$.

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

$$Z_i(s) = \omega Z_M(s) + \sqrt{1-\omega^2} W_i(s), \qquad i = 1, \ldots, 11. \tag{3.7}$$

The third step implies that each field in the ensemble is a Gaussian process with mean zero, variance one, range a_M, smoothness ν_M, and univariate "forecast skill" controlled by the parameter ω (see Appendix A.2). 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 A.3), and thus the ensemble forecasts are calibrated in the univariate sense.

In this study, fields were constructed on a square grid $[-20, 20]$ with resolution 0.2. Verification-ensemble samples were collected by repeating the simulation above five-thousand 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 total of seventy-seven parameters sets investigated. 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 determined by the time period for which the verification is performed. For 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.

3.3 Simulation analysis

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

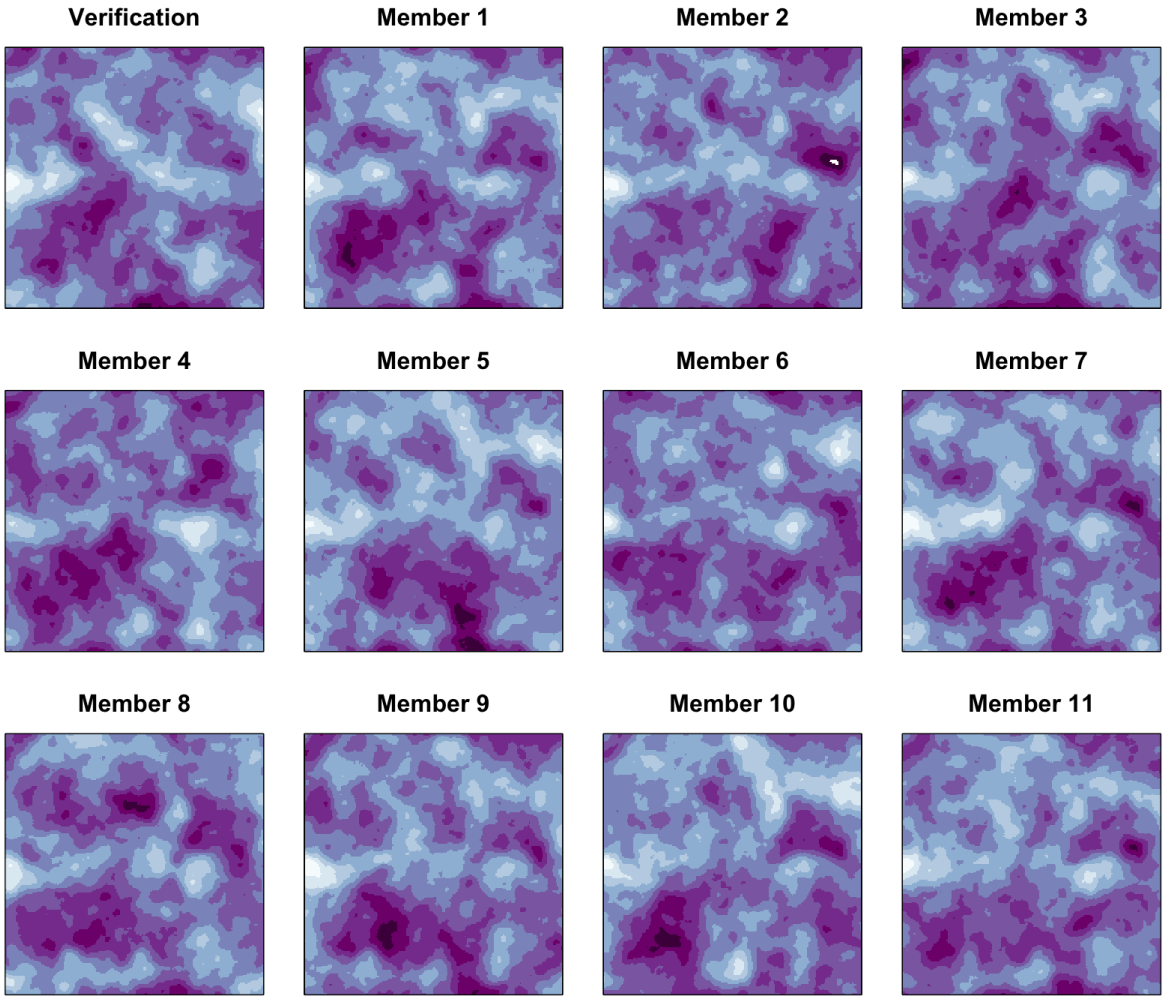

Figure 3.1: Simulated verification field and a corresponding ensemble (arbitrary color scale). In this example, the verification has correlation length $a_0 = 2$, and the spatial structure of the ensemble is calibrated in the sense that $a_M = a_0$.

3.3.1 Illustrative examples of FTE histograms

First, we study the discrimination ability of the FTE in something of an exaggerated setting, where the miscalibration is obvious. We choose the mean 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. 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. 3.2, row A), correct correlation length using ratio 1.0 (Fig. 3.2, row B), and too large

of a correlation length using ratio 1.5 (Fig. 3.2, row C). Corresponding FTE histograms are then constructed with respect to these three ratios using five-thousand verification-ensemble samples in each case. This revealing example is depicted in Fig. 3.2 and behaves as described in Table 2.1, where the FTE histogram takes a ∪-shape when the ensemble correlation length is too small, a ∩-shape when the correlation length is too large, and is approximately flat when the ensemble fields have the same correlation length as the verification field.

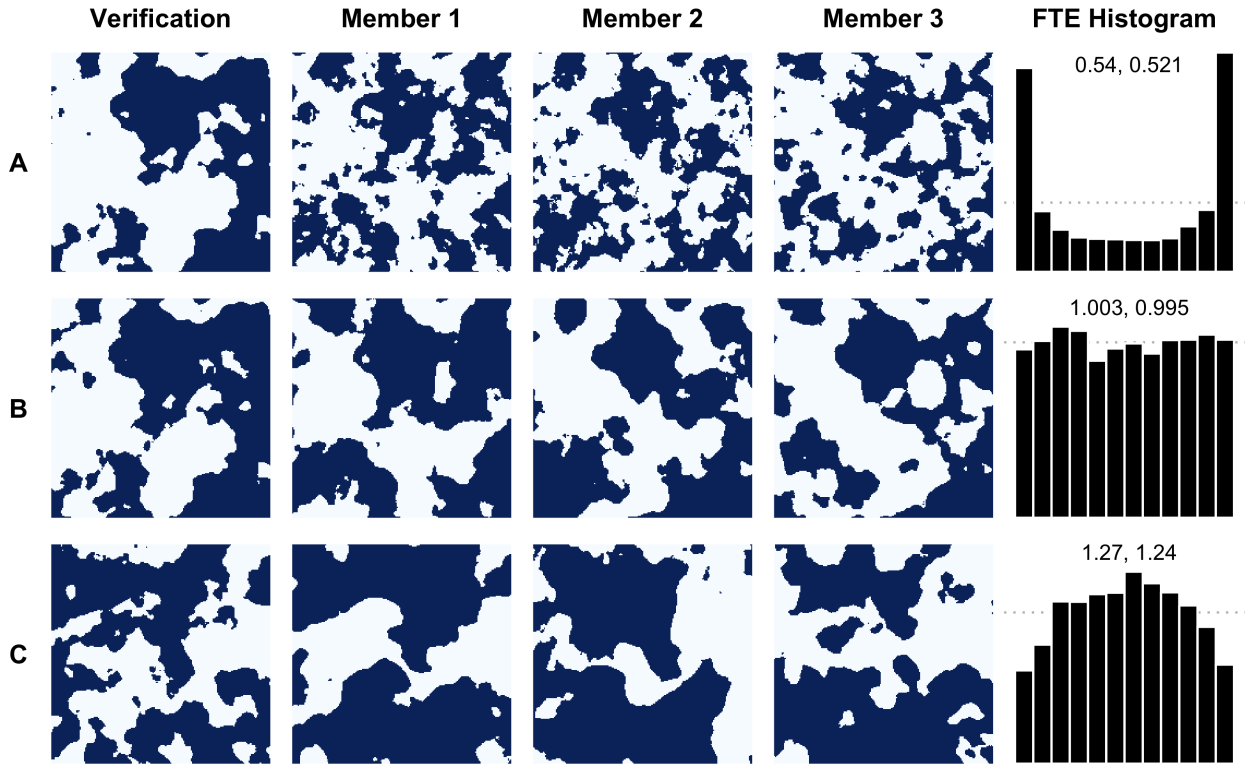

Figure 3.2: Example binary exceedance verification field and a subset of ensemble fields with representative FTE histogram for threshold $\tau = 0$ found using 5000 samples. Dark blue regions indicate threshold exceedance. All verification fields have correlation length $a_0 = 2$ and ensemble fields have correlation length $a_M = 1, 2, 3$ in rows A, B, and C respectively. FTE histograms are density histograms with a dotted line $y = 1$ and estimated beta distribution parameters annotated.

While the FTE is able to correctly identify the obvious miscalibration of the ensemble for the scenario in Fig. 3.2, one could likely draw the same conclusions by visual inspection and would not use the FTE 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 is realized when the miscalibration is not so visually obvious. This more realistic example is illustrated in Fig. 3.3 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 is noticeably ∩-shaped. Again, these results are consistent with Table 2.1, and we conclude that the FTE metric maintains accurate discrimination ability even when ensemble members are only slightly miscalibrated.

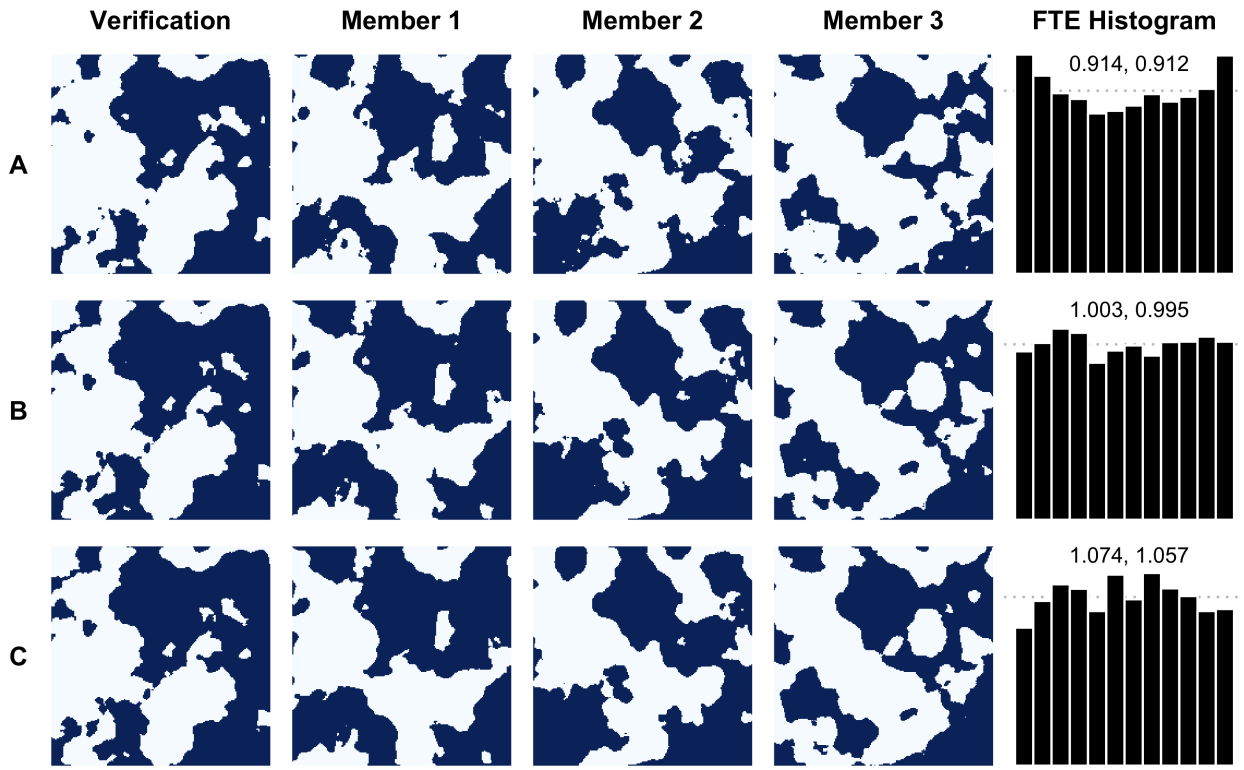

Figure 3.3: Same as Fig. 3.2, but ensemble fields have ranges $a_M = 1.8, 2, 2.2$ in rows A, B, and C respectively.

Of course, one may often want to use a threshold parameter other than the mean of the marginal distribution. The choice of τ is somewhat application specific; for example, it can be

Figure 3.4: Same as Fig. 3.3, but for threshold $\tau = 2$.

chosen such as to focus on high precipitation amounts. Thus, it is important that the FTE metric maintains discrimination ability for different choices of τ. For a visual example, the same experiment depicted in Fig. 3.3 is repeated in Fig. 3.4, 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 $\cup$-shaped and has a slight right-skew. When the ensemble correlation length is slightly too large (row C), the FTE histogram is $\cap$-shaped and 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 2.1, the effect of the threshold can be studied more generally using the estimated beta parameters.

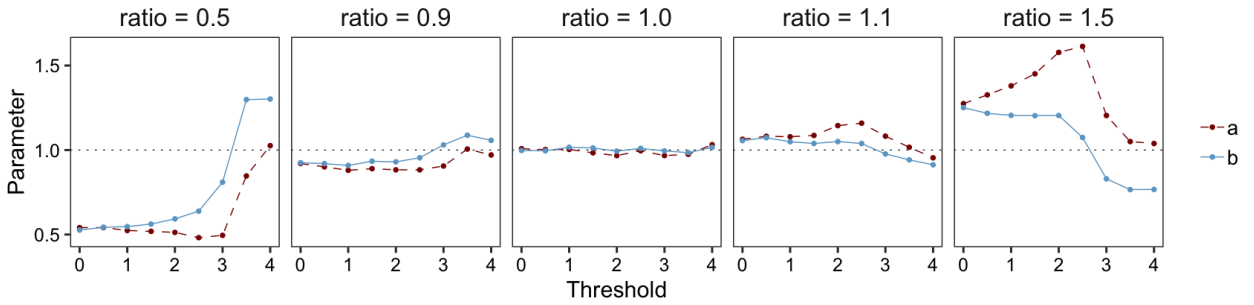

Figure 3.5: Estimated beta distribution parameters of FTE histograms calculated over different thresholds for forecasts with low, even, and high correlation length ratios about $a_0 = 2$.

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 $a = b = 1$, the FTE histogram is perfectly uniform. How do these parameter values change as the threshold τ increases? Figure 3.5 shows that when the correlation length ratio is 1.0, the estimated beta distribution parameters are both about 1 for every choice of τ, correctly indicating flat histograms. When the correlation length ratio is less than 1.0, the beta parameters are themselves generally less than 1, indicating $\cup$-shaped histograms. Conversely, when the ratio is greater than 1.0, the beta parameters are themselves greater than 1, indicating $\cap$-shaped histograms. There is an interesting tendency toward unity in the beta parameters for high thresholds (i.e., τ greater than three standard deviations from the mean). This is because the number of exceedances for high thresholds will often be zero. For thresholds greater than two standard deviations from the mean, the percentage of ranks discarded from the histogram becomes increasingly significant. Moreover, since ranks are only discarded from the histogram if all ensemble members and the verification field have the same FTE, the FTE histogram for high thresholds will be composed largely of ranks resulting from ties broken uniformly at random, i.e. there are only a few unique values in each sample of π.

Illustrating this phenomenon, Figs. 3.6 and 3.7 complement Fig. 3.5 for correlation length ratios of 0.5 and 1.1, respectively. Note that the shape of the FTE histograms (described by their

estimated beta parameters) consistently captures the type of miscalibration for $\tau < 3$. Once $\tau = 3$, artificial uniformity is introduced and the FTE histogram becomes skewed. For $\tau = 4$ (a threshold four standard deviations from the mean), the FTE histogram is calculated using a low percentage of samples which are themselves largely based ties broken uniformly at random. This results in a deceptively uniform histogram which also suffers from skew leading to estimated beta parameters that are close to 1 but ultimately mislabel the type of miscalibration in the ensemble. Therefore, when working with very high thresholds (e.g. extreme value analysis) we advise the practitioner to consider the percentage of samples used in the calculation of the FTE histogram with the largest sample size available.

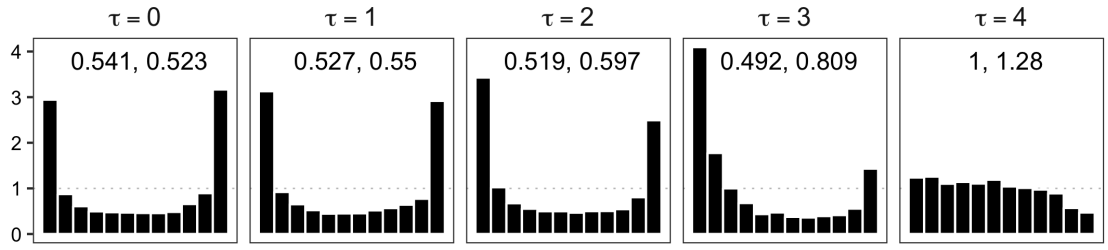

Figure 3.6: FTE histograms calculated over different thresholds for a correlation length ratio of 0.5 about $a_0 = 2$. At $\tau = 3$, no samples are discarded. With $\tau = 4$, 1996 ranks are discarded (nearly 40%).

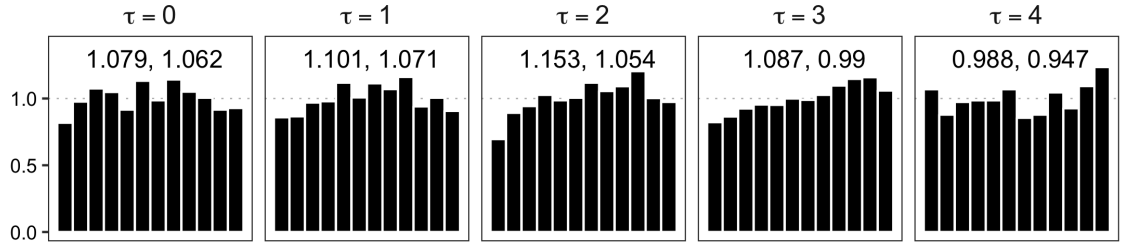

Figure 3.7: Same as Fig. 3.6, but for a correlation length ratio of 1.1. With $\tau = 3$, 187 ranks are discarded (about 4%). At $\tau = 4$, 3990 ranks are discarded (nearly 80%).

Another variable of interest in evaluating the FTE metric 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" and larger correlation length mimics a "small domain." Analyzing estimated beta parameters over a range of correlation length ratios is then equivalent to studying the FTE metric's 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. In either case, Fig. 3.8 shows that the beta parameters estimated using the FTE metric quickly rise above or fall below unity when the correlation length ratio is not 1.0 itself. The steep slopes around the correlation length of 1.0 in both cases indicate that the FTE metric maintains good discrimination ability regardless of domain size. Figure 3.9 complements Fig. 3.8 for the case of $a_0 = 1$, illustrating how the shape of the FTE histograms changes as the correlation length ratio varies.

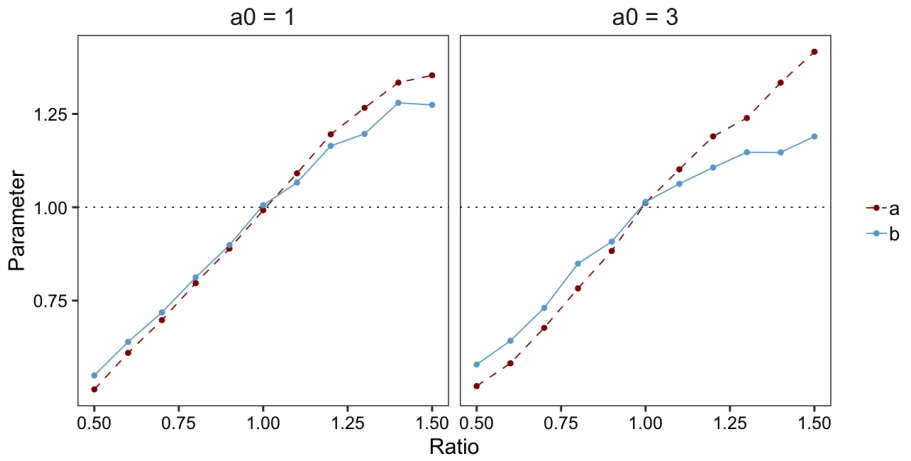

Figure 3.8: Estimated beta parameters of FTE histograms for verification correlation length $a_0 = 1, 3$ and ensemble range a_M varying from $0.5a_0$ to $1.5a_0$, with $\tau = 1$.

We now turn attention back to our motivating figure (Fig. 1.1), 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.

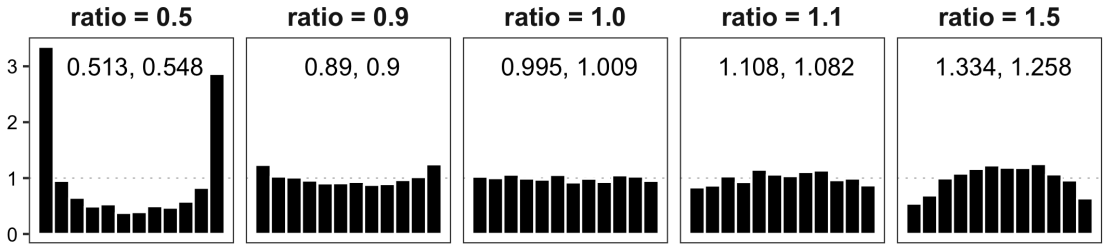

Figure 3.9: FTE histograms calculated for different correlation length ratios about $a_0 = 1$ using threshold $\tau = 1$.

However, as demonstrated by the analysis above (Fig. 3.2 – Fig. 3.4), these misspecifications are certainly identifiable using the FTE metric.

Chapter 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 numerical weather prediction 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 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 United States 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 ∼0.5°. Downscaling and verification is performed against precipitation analyses from the ∼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. 4.1 for a visual reference.

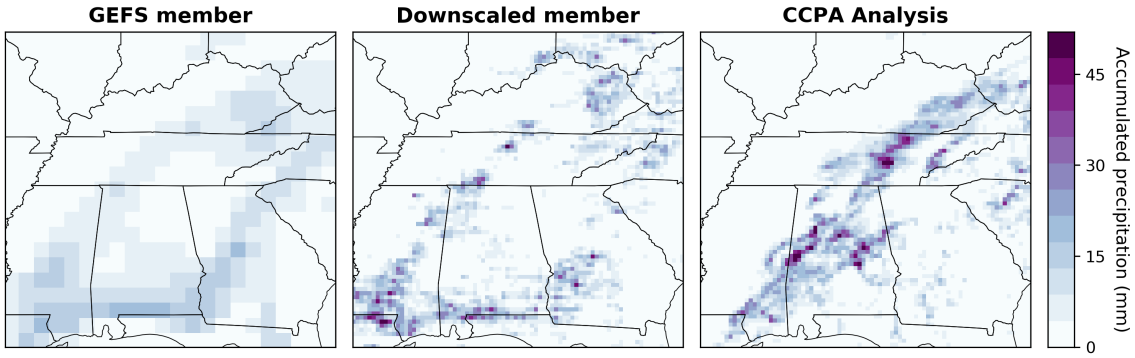

Figure 4.1: Examples of different data fields for 6-hour precipitation accumulation on July 24, 2004. From left to right: coarse-scale GEFS ensemble member, the same member downscaled to the analysis resolution, and the corresponding CCPA analysis.

4.2 Univariate verification

Before applying the FTE 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, 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. 4.2. Note that grid points receiving the same rank for the observation and all forecast fields – for example, when there is zero accumulation at a point for all fields – are withheld from the histogram to avoid artificial uniformity introduced by breaking ties in rank at random.

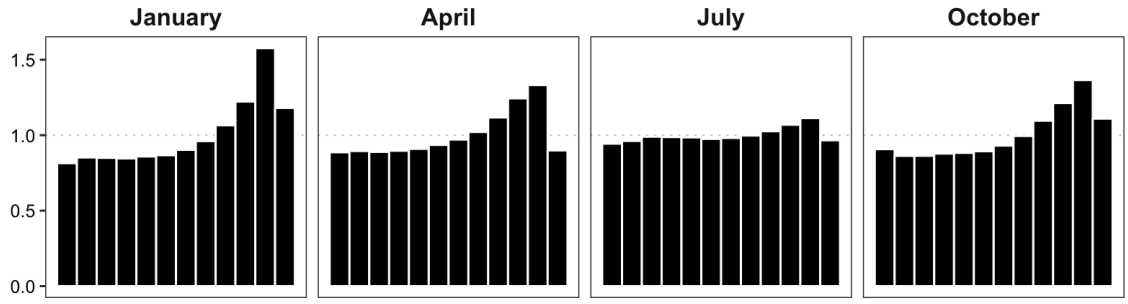

Figure 4.2: Verification rank histograms (density) for downscaled fields at representative months with cases of fully tied ranks removed.

Recall that a calibrated forecast should yield rank histograms that are approximately uniform. Clearly, the rank histograms for the downscaled forecast fields shown in Fig. 4.2 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 with our analysis of the spatial calibration of the downscaled ensemble forecasts fields, keeping in mind though that some of the possible non-uniformity of FTE histograms in January, April and October could be due to univariate miscalibration.

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 relatively flat and 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. 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. 4.3, it is clear by visual inspection that the FTE histograms are all ∪-shaped to some extent, though the fitted beta parameters 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

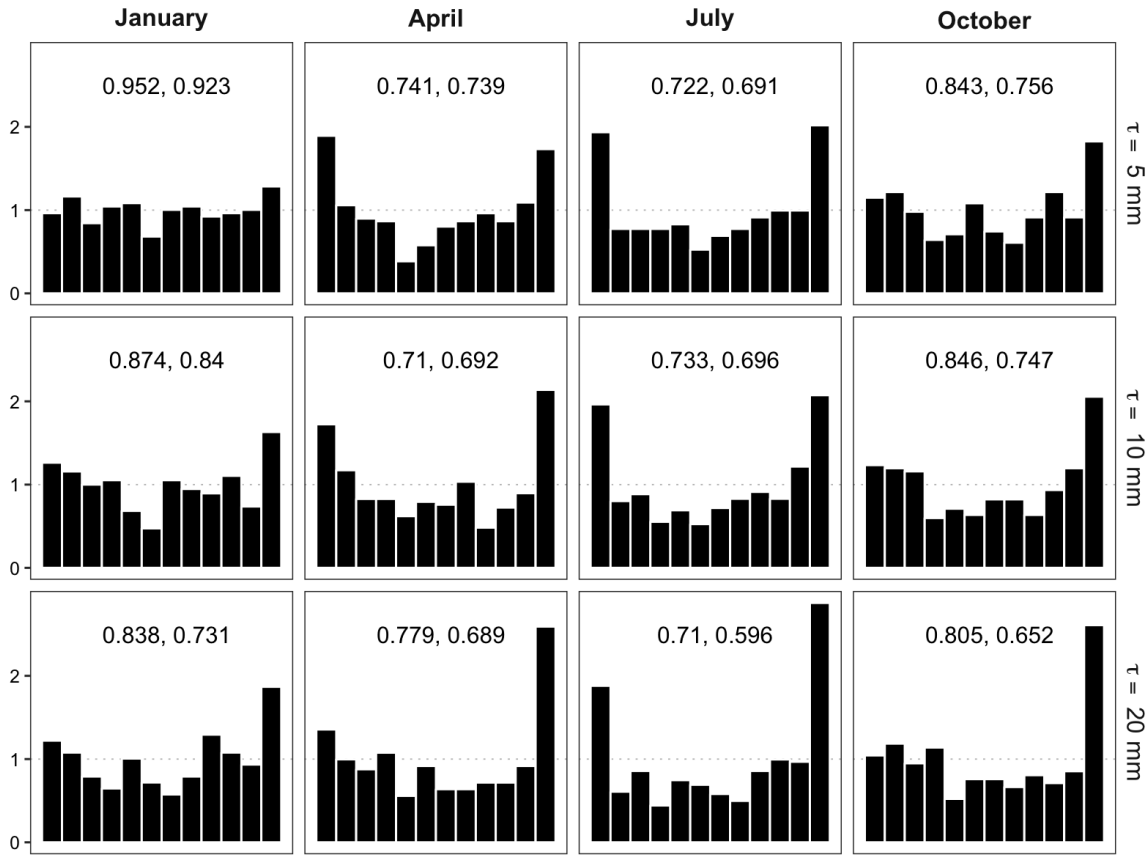

Figure 4.3: FTE histograms for downscaled fields at different thresholds in representative months with estimated beta distribution parameters.

would suggest that the downscaled ensemble overestimates fine scale variability during the seasons with more convective 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 can thus provide valuable diagnostic information that helps identify shortcomings of a forecast methodology. Indeed, current research seeks 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).

Chapter 5

Conclusion

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 thesis, 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 stochastically disaggregated FTE histogram by maximum likelihood to summarize 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 can be compared according to Table 2.1 to obtain a simple objective summary of dispersion and bias in the ensemble forecast. 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.

Bibliography

Anderson, J. L. (1996). A Method for Producing and Evaluating Probabilistic Forecasts from Ensemble Model Integrations. Journal of Climate, 9(7):1518–1530.

Apanasovich, T. V., Genton, M. G., and Sun, Y. (2012). 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.

Bellier, J., Scheuerer, M., and Hamill, T. M. (2020). Advances in statistical precipitation downscaling using gibbs sampling. To be submitted to Journal of Hydrometeorology.

Buizza, R., Bidlot, J.-R., Wedi, N., Fuentes, M., Hamrud, M., Holt, G., and Vitart, F. (2007). The new ECMWF VAREPS (variable resolution ensemble prediction system). Quarterly Journal of the Royal Meteorological Society, 133(624):681–695.

Buschow, S., Pidstrigach, J., and Friederichs, P. (2019). Assessment of wavelet-based spatial verification by means of a stochastic precipitation model (wv_verif v0.1.0). Geoscientific Model Development, 12(8):3401–3418.

Feldmann, K., Scheuerer, M., and Thorarinsdottir, T. L. (2015). Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous gaussian regression. Monthly Weather Review, 143(3):955–971.

Gagnon, P., Rousseau, A. N., Mailhot, A., and Caya, D. (2012). Spatial disaggregation of mean areal rainfall using gibbs sampling. Journal of Hydrometeorology, 13(1):324–337.

Genton, M. G. and Kleiber, W. (2015). Cross-covariance functions for multivariate geostatistics. Statistical Science, 30:147–163.

Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., and Ebert, E. E. (2009). Intercomparison of spatial forecast verification methods. Weather and Forecasting, 24(5):1416–1430.

Gneiting, T., Kleiber, W., and Schlather, M. (2010). Matérn Cross-Covariance Functions for Multivariate Random Fields. Journal of the American Statistical Association, 105(491):1167–1177.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., and Johnson, N. A. (2008). Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. TEST, 17(2):211.

Hamill, T. M. (2001). Interpretation of Rank Histograms for Verifying Ensemble Forecasts. Monthly Weather Review, 129(3):550–560.

Hamill, T. M., Bates, G. T., Whitaker, J. S., Murray, D. R., Fiorino, M., Galarneau, T. J., Zhu, Y., and Lapenta, W. (2013). NOAA's second-generation global medium-range ensemble reforecast dataset. Bulletin of the American Meteorological Society, 94(10):1553–1565.

Hou, D., Charles, M., Luo, Y., Toth, Z., Zhu, Y., Krzysztofowicz, R., Lin, Y., Xie, P., Seo, D.-J., Pena, M., and Cui, B. (2014). Climatology-calibrated precipitation analysis at fine scales: Statistical adjustment of stage IV toward CPC gauge-based analysis. Journal of Hydrometeorology, 15(6):2542–2557.

Keller, J. D. and Hense, A. (2011). A new non-gaussian evaluation method for ensemble forecasts based on analysis rank histograms. Meteorologische Zeitschrift, 20(2):107–117.

Kleiber, W. (2017). Coherence for multivariate random fields. Statistica Sinica, 27:1675–1697.

Leutbecher, M. and Palmer, T. (2008). Ensemble forecasting. Journal of Computational Physics, 227(7):3515 – 3539.

Murphy, A. H. and Winkler, R. L. (1977). Reliability of subjective probability forecasts of precipitation and temperature. Journal of the Royal Statistical Society. Series C (Applied Statistics), 26(1):41–47.

Schefzik, R., Thorarinsdottir, T. L., and Gneiting, T. (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. Statistical Science, 28(4):616–640.

Scheuerer, M. and Hamill, T. M. (2015). Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. Monthly Weather Review, 143(11):4578–4596.

Scheuerer, M. and Hamill, T. M. (2018). Generating Calibrated Ensembles of Physically Realistic, High-Resolution Precipitation Forecast Fields Based on GEFS Model Output. Journal of Hydrometeorology, 19(10):1651–1670.

Schlather, M., Malinowski, A., Menck, P., Oesting, M., and Strokorb, K. (2015). Analysis, simulation and prediction of multivariate random fields with package RandomFields. Journal of Statistical Software, Articles, 63(8):1–25.

Smith, L. A. and Hansen, J. A. (2004). Extending the limits of ensemble forecast verification with the minimum spanning tree. Monthly Weather Review, 132(6):1522–1528.

Thorarinsdottir, T. L., Scheuerer, M., and Heinz, C. (2016). Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. Journal of Computational and Graphical Statistics, 25(1):105–122.

Toth, Z. and Kalnay, E. (1993). Ensemble forecasting at nmc: The generation of perturbations. Bulletin of the American Meteorological Society, 74(12):2317–2330.

Wilks, D. S. (2004). The minimum spanning tree histogram as a verification tool for multidimensional ensemble forecasts. Monthly Weather Review, 132(6):1329–1340.

Zhou, X., Zhu, Y., Hou, D., Luo, Y., Peng, J., and Wobus, R. (2017). Performance of the new NCEP global ensemble forecast system in a parallel experiment. Weather and Forecasting, 32(5):1989–2004.

Ziegel, J. F. and Gneiting, T. (2014). Copula calibration. Electronic Journal of Statistics, 8(2):2619–2638.

Appendix A

Statistical properties and methods

A.1 Stochastic disaggregation of transformed ranks

The vector of ranks $\mathbf{r}$ 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

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

(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].

A.2 Properties of simulated ensemble members

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

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

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

$$\begin{aligned} \text{Var}[Z_i] &= \text{Var}[\omega Z_M + \sqrt{1-\omega^2} W_i] \\ &= \omega^2 \text{Var}[Z_M] + (1-\omega^2)\text{Var}[W_i] \\ &= 1 \end{aligned} \tag{A.2}$$

using independence of 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

$$\begin{aligned} &\text{Cov}[Z_i(s+h), Z_i(s)] \\ &\quad = \text{Cov}[\omega Z_M(s+h) + \sqrt{1-\omega^2} W_i(s+h), \omega Z_M(s) + \sqrt{1-\omega^2} W_i(s)] \\ &\quad = \omega^2 \text{Cov}[Z_M(s+h), Z_M(s)] + \omega\sqrt{1-\omega^2}\text{Cov}[Z_M(s+h), W_i(s)] \\ &\qquad + \omega\sqrt{1-\omega^2}\text{Cov}[W_i(s+h), Z_M(s)] + (1-\omega^2)\text{Cov}[W_i(s+h), W_i(s)] \\ &\quad = \omega^2 \text{Cov}[Z_M(s+h), Z_M(s)] + (1-\omega^2)\text{Cov}[W_i(s+h), W_i(s)] \\ &\quad = \omega^2 M(h|\nu_M, a_M) + (1-\omega^2) M(h|\nu_M, a_M) \\ &\quad = M(h|\nu_M, a_M) \end{aligned} \tag{A.3}$$

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 $Z_M(s)$.

A.3 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

$$Z_i = \omega Z_M + \sqrt{1-\omega^2} W_i, \quad i = 1, \ldots, k, \quad \omega \in [-1, 1]. \tag{A.4}$$

From Appendix A.2 we have that each of Z_i is again a standard Gaussian random variable. Then, for $i \neq j$, we see that

$$\begin{aligned}
\mathrm{Cov}[Z_i, Z_j] &= \mathbb{E}[(Z_i - \mathbb{E}[Z_i])(Z_j - \mathbb{E}[Z_j])] \\
&= \mathbb{E}[Z_i Z_j] \\
&= \mathbb{E}[(\omega Z_M + \sqrt{1-\omega^2} W_i)(\omega Z_M + \sqrt{1-\omega^2} W_j)] \\
&= \mathbb{E}[\omega^2 Z_M^2 + \omega\sqrt{1-\omega^2} Z_M (W_i + W_j) + (1-\omega^2) W_i W_j] \\
&= \omega^2 \mathrm{Var}[Z_M] + (1-\omega^2)\mathrm{Cov}[W_i, W_j] \\
&= \omega^2.
\end{aligned} \tag{A.5}$$

Using a similar technique we see

$$\begin{aligned}
\mathrm{Cov}[Z_0, Z_i] &= \mathbb{E}[(Z_0 - \mathbb{E}[Z_0])(Z_i - \mathbb{E}[Z_i])] \\
&= \mathbb{E}[Z_0 Z_i] \\
&= \mathbb{E}[Z_0(\omega Z_M + \sqrt{1-\omega^2} W_i)] \\
&= \mathbb{E}[\omega Z_0 Z_M] + \mathbb{E}[\sqrt{1-\omega^2} Z_0 W_i] \\
&= \omega \mathrm{Cov}[Z_0, Z_M] \\
&= \omega\rho.
\end{aligned} \tag{A.6}$$

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

$$\mathrm{Cov}[\mathbf{Z}] = \begin{pmatrix}
1 & \omega\rho & \ldots & \ldots & \omega\rho \\
\omega\rho & \ddots & \omega^2 & \ldots & \omega^2 \\
\vdots & \omega^2 & \ddots & \ddots & \vdots \\
\vdots & \vdots & \ddots & \ddots & \omega^2 \\
\omega\rho & \omega^2 & \ldots & \omega^2 & 1
\end{pmatrix} \tag{A.7}$$

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.