NOTES AND CORRESPONDENCE

Extending the Limits of Ensemble Forecast Verification with the Minimum Spanning Tree

LEONARD A. SMITH

Department of Mathematics, University of Oxford, Oxford, United Kingdom

JAMES A. HANSEN

Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts

3 March 2003 and 7 January 2004

ABSTRACT

Uncertainty in the initial condition is one of the factors that limits the utility of single-model-run predictions of even deterministic nonlinear systems. In practice, an ensemble of initial conditions is often used to generate forecasts with the dual aims of 1) estimating the reliability of the forecasts and 2) estimating the probability distribution of the future state of the system. Current rank histogram ensemble verification techniques can only evaluate scalars drawn from ensembles and associated verification; a new method is presented that allows verification in high-dimensional spaces, including those of the verifications for 10^6 dimensional numerical weather prediction forecasts.

As has long been recognized (Thompson 1957), uncertainty in the initial condition limits the utility of a single model run as a forecast of a nonlinear system like the earth's atmosphere. If this uncertainty is accepted, then internal consistency requires that an ensemble of initial conditions, each consistent with the observations, be evolved forward under the model. This note considers not the selection of initial conditions but rather the evaluation of subsequent forecasts; it presents a new method of substantially greater applicability than current methods. We will focus on the property of *reliability*: the property that when a given probability distribution P has been forecast, the conditional probability distribution of the verification is equal to P. Methods for evaluating ensemble reliability (Murphy and Winkler 1987) include those based on rank order statistics; current implementations are restricted to a single scalar variable drawn from the forecast state. A generalization of this approach to higher-dimensional spaces will have wide application in both the evaluation of physical models and in forecasting dynamical systems ranging from low-dimensional chaotic systems to high-dimensional weather models.

Corresponding author address: Dr. James A. Hansen, Dept. of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139.
E-mail: jhansen@mit.edu

Ensemble forecasting replaces a single ''best first guess'' initial condition with a relatively small ensemble of initial conditions, each consistent with the observational uncertainty in the initial state of the system. Closure problems (Epstein 1969; Leith 1974) prevent an analytic solution, motivating this Monte Carlo approach. Methods for selecting these initial conditions date back to Lorenz (1965), and competing operational approaches, evolved from the early 1990s, are used in the European, American, and Canadian weather forecasting centers (Molteni et al. 1996; Toth and Kalnay 1993; Houtekamer et al. 1996). By observing how quickly the ensemble spreads out (or shrinks), one obtains a local estimate of the stability of forecasts made in this region of state space. Global measures like Lyapunov exponents are useless here (Smith et al. 1999), except in the most simple, uniform systems. Even localized Lyapunov exponents are misleading (Ziehmann et al. 2000), since they are based on the linearized dynamics of infinitesimals, while the ensemble samples the relevant nonlinearities. Indeed, chaos places no a priori limits on predictability: given a perfect model, ensembles exist that will slowly evolve toward the invariant measure of the system, but the time scale on which this happens is independent of the measures used to define chaos, which are in turn based upon the statistics of infinitesimals (Oseledec 1968; Smith et al. 1999).

The basic difficulty in evaluating ensemble forecasts

stems from the fact that the ensemble forecast represents a draw from an estimate of a probability density function (PDF) of forecasts in the model state space (Palmer 2000), while the verification (the true state of the system at the forecast time or its observation, depending upon what is available) is a single state.[1] Further, since each forecast is evolved from a different initial state (van den Dool 1994), the details of the forecast PDF differ for each forecast.

A standard ensemble reliability evaluation method is based on the rank order statistic of the forecast (Anderson 1996; Talagrand et al. 1999; Hamill 2001); after sketching this method, a generalization to higher-dimensional systems based on minimum spanning trees is presented. The new method is illustrated in a 2-dimensional and a 69 173-dimensional example.

Rank-order-statistic-based methods aim to assess whether ensemble members and the verification are drawn from the same probability density function. If this is the case then no statistic can distinguish the verification from the ensemble members. In particular, if n scalar forecast values and a (single) verification are arranged in order of increasing magnitude, the verification is equally likely to take each position in this rank ordering (von Mises 1981), and therefore, the number of ensemble members smaller than the verification will be uniformly distributed[2] between zero and n; on average, the verification is equally likely to fall into any of the $n + 1$ ''bins'' defined by the rank-ordered ensemble members. The shape of the forecast PDF changes with each forecast, but for each forecast the probability of having exactly i ensemble members smaller than the verification is $1/(n + 1)$ for each value of i in (0, 1, 2, 3, . . . , n); hence a histogram constructed by collecting the rank order of verification over N realizations should be flat: the expected fraction of members in a bin having mean $1/(n + 1)$ and standard deviation $(1/\sqrt{N})\sqrt{1/(n + 1)(1 - 1/(n + 1))}$.

The strength of this approach is that it permits the examination (and hopefully improvement) of forecasts of a given variable at a given location, say, 1-day forecasts of the temperature in London, United Kingdom. One cannot, however, evaluate the diagrams from more than one variable unless the forecast value of each variable is truly independent. While in the long run one would still hope for each bin to have mean $1/(n + 1)$, there is no general method for assessing what counts as flat: not knowing the expected standard deviation in the multivariable case makes it impossible to evaluate the significance of the results in general, although bootstrap resampling techniques (Efron and Tibshirani 1993) provide a test for necessary conditions.[3] In this work we follow the common practice of calling rank histograms based on a single scalar value drawn from each ensemble member and verification ''Talagrand diagrams.''

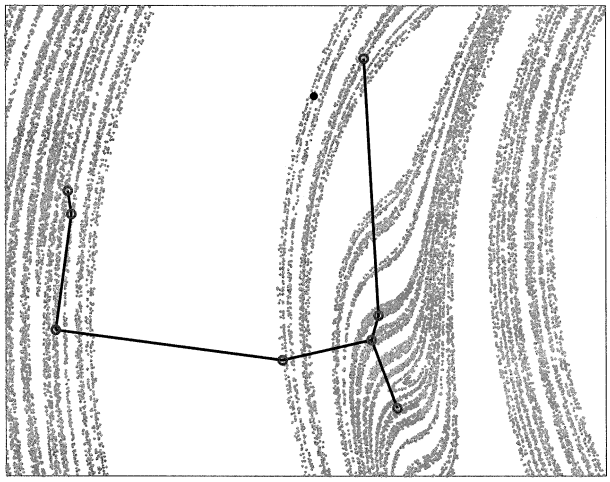

FIG. 1. An example of a minimum spanning tree. The small dots are points on the Ikeda attractor. The open circles are ensemble members, and the lines connecting them indicate the MST for the L2 norm. The large solid dot is the verification.

What is needed is a generalization of the rank ordering of a single variable to the higher-dimensional case (Smith 2000). This letter presents a new approach based on minimum spanning trees. Given a set of n points in an m-dimensional space, a spanning tree is a collection of $n - 1$ pairs of points (branches) such that all points are used at least once. Defining a metric on the space[4] associates with each tree a length, specifically the sum of the distances of each branch in the tree. The spanning tree with the smallest length is the minimum spanning tree (MST). An example is shown in Fig. 1. MSTs are well known in graph theory and are commonly used as a method for classification and for network design. The classic algorithms for constructing MSTs are found in Kruskal (1956) and Prim (1957).

This MST approach yields rank histograms that can evaluate ensembles in an m-dimensional space. When assessing a forecast we have $n + 1$ points (n ensemble members and the verification); let l_0 be the length of the MST constructed using only the ensemble members (e.g., the MST resulting from only the eight open circles in Fig. 1), and l_i be the length of the MST where the verification is used in place of the ith ensemble member. Again, if the verification is drawn from the same distribution as the ensemble members then the number of times that precisely i trees have length smaller than l_0

[1] The authors agree with an anonymous reviewer that ideally one should assess in observation space (Smith 2003); in practice, verification in model space is common.

[2] The sense of this definition (smaller than the verification versus larger than the verification) is opposite to that in Smith (2000).

[3] In short, while it is in general never possible to say whether such a histogram is flat, the bootstrap may allow us to say that it is not flat.

[4] The choice of this metric, or norm, is of critical importance and will be problem dependent.

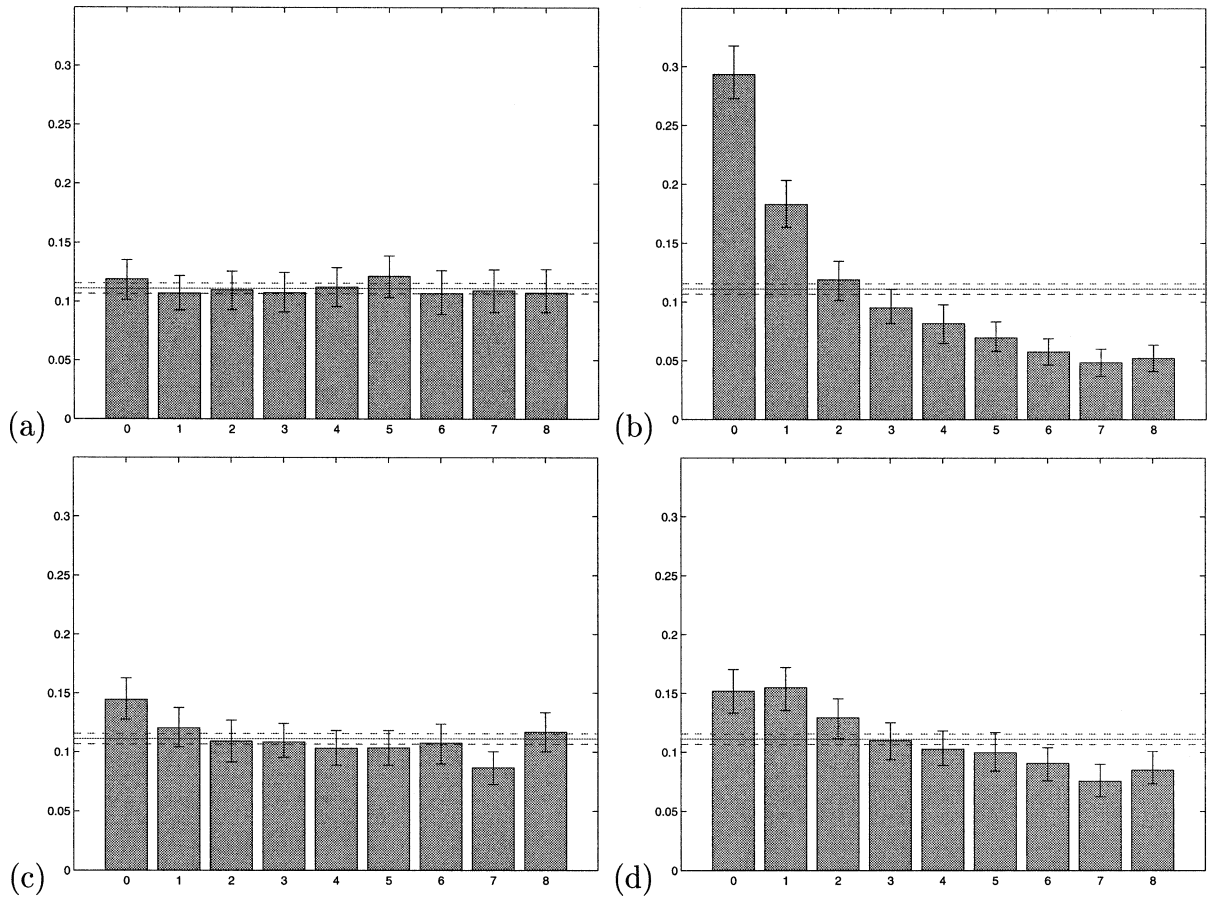

FIG. 2. Minimum spanning tree rank histograms. Ensemble members are always drawn from the Ikeda attractor, while verification differs for each panel: (a) verification is also drawn from the attractor, (b) verification is drawn randomly from a box in the area of interest, (c) verification is drawn randomly from a line that is the best linear fit to the local Ikeda attractor structure, and (d) the x component and the y component of the verification are drawn *independently* from the Ikeda attractor distribution. The solid horizontal line is the expected mean, and the horizontal dashed lines are the expected 1 std dev bounds. The vertical lines at the top of the bar in each bin are produced by bootstrapping (resampling with replacement) from the data that was used to construct the rank histograms. They represent the 99% bound on expected values.

is a random variable with mean $1/(n + 1)$ and standard deviation $(1/\sqrt{N})\sqrt{1/(n + 1)(1 - 1/(n + 1))}$. Suppose, for example, that the verification tends to be far from the ensemble: in this case l_0 will tend to be *smaller* than the l_i and the histogram will reflect a systematic bias (excessive population in the bins of small i). It is important to point out that the MST rank histogram will reflect projection effects. Projecting (assimilating) verifying observations into model space can result in verifications that are artificially close to ensemble members. Projecting ensemble members into observation space can have the corresponding effect.

The behavior of the MST rank histogram method is demonstrated using the Ikeda (1979) system. Four different scenarios are considered. In each scenario ensembles of size $n = 8$ are drawn from the correct Ikeda distribution,[5] while four different methods are used to produce the verification. Statistics are collected over $N = 5000$ ensemble realizations and results are displayed in Fig. 2. Figure 2a shows the MST rank histogram produced when the verifications, like the ensembles, are drawn from the Ikeda distribution. For each of the N realizations, a location on the Ikeda attractor is chosen at random, and ensemble members and verification are chosen from points on the attractor that are within an ϵ-sided box centered at that location. For the results shown here, $\epsilon = 10\%$ the size of the Ikeda attractor. In this scenario it is expected that each histogram bin will have a value of $1/(n + 1) = 1/9$ (given by the solid horizontal line) and a standard deviation of $(1/\sqrt{N})\sqrt{1/(n + 1)(1 - 1/(n + 1))} = (1/\sqrt{5000})\sqrt{(1/9)(1 - 1/9)}$ (given by the dashed horizontal lines), values that are consistent with the results shown in Fig. 2a. The vertical lines centered at the top of the bar in each bin are the result of a bootstrap uncertainty estimate (Efron and Tibshirani 1993). They

[5] Each ensemble member lies on the Ikeda attractor.

are produced by resampling (with replacement) from the N realizations used to generate the MST rank histograms and recording the MST rank histograms that result for a number of different resamplings. The bars represent the nonparametric 99% confidence bound on expected values. The extent to which the bars do not overlap the expected mean (the solid horizontal line) provides information about the extent to which the histogram is not flat.

Figure 2b shows the MST rank histogram for the case where verification is drawn *randomly* from within each ϵ box. Because points on the Ikeda attractor are not uniformly distributed in such a box, verification will, on average, lie far from any generated ensemble and one would thus expect l_0 values to tend to lie below l_i values, as is seen in Fig. 2b. In Fig. 2c verification is restricted to lie on a line that represents a linear fit to points lying on the Ikeda attractor in the domain of interest. Again the MST rank histogram reveals the fact that the verification is not drawn from the same distribution as the ensemble: the MST rank histogram is not flat, with bootstrap results for two of the bins having no overlap with the expected mean. Finally, in Fig. 2d results are shown for the case when verification's x component is drawn from the correct Ikeda x distribution and the verification's y component is drawn from the correct Ikeda y distribution but is independent of the x component selected. The componentwise distributions used to construct the ensembles are the same as those used to construct the verifications, so Talagrand diagrams constructed for each component would each yield uniform distributions. But the histogram in Fig. 2d is not uniform, again highlighting the strength of the multidimensional MST rank histogram approach. The MST rank histogram is able to discern that the *two-dimensional distribution* used to construct the verification differs from the two-dimensional distribution used to construct the ensembles. The nonuniform nature of Figs. 2b–d is accentuated for larger ensemble sizes (results not shown).

MST rank histograms also suggest a simple method for comparing the quality of a collection of ensemble prediction systems (EPSs). For each EPS, consider the distribution of the bin populations, in particular the distance of the relative frequency of each bin minus the expected value, measured in terms of the number of standard deviations it is from the expected value. Plotting the cumulative distribution function (CDF) of this distribution for each EPS roughly indicates how the EPSs compare: methods that fall consistently to the left (close to the expected values) are preferred. The approach is illustrated in Fig. 3 for the EPSs used in Fig. 2. The thick black line is for verification drawn from the Ikeda attractor (Fig. 2a), the thin black line is for verification drawn randomly from a box (Fig. 2b), the thick dashed line is for verification drawn from a line (Fig. 2c), and the thick dash–dot line is for verification drawn independently from the correct x and y distributions (Fig. 2d). The thin gray lines indicate the results from 1000 realizations of random draws from a known Gaussian distribution. Note that the rank ordering of the EPSs is for verifications drawn 1) from the attractor, 2) from a line, 3) from x and y independently from the correct distributions, and 4) randomly.

Simply summing the rank histograms from many grid points (e.g., 500-hPa height at a number of locations around the Northern Hemisphere) yields a histogram that is difficult to interpret. It is tempting to try to extend the Talagrand diagram concept to multiple dimensions by empirically identifying independent state space directions [e.g., by using empirical orthogonal functions (EOFs)] and summing the Talagrand diagrams that result from projecting onto these directions. But in nonlinear systems EOFs are not independent, and Talagrand diagrams summed over EOFs suffer from the same limitations as Talagrand diagrams summed over model grid points. In the meteorological and oceanographic context, relevant EOF patterns might include the Arctic Oscillation pattern, the El Niño–Southern Oscillation pattern, and the Pacific–North America pattern. Assessing the Talagrand diagram for any one of these patterns is well defined, but the interdependence of the physical processes that drive these patterns makes it unclear how to interpret the collection of Talagrand diagrams defined by each pattern. The MST rank histogram does not suffer from these limitations. An MST can be constructed in EOF space just as easily as in model space. And one need not be limited to analysis in EOF space. Any scale decomposition can be used to address questions like, At what scale do the ensemble forecasts appear reliable?

An MST diagram is easily constructed for the ensemble output from operational NWP centers. Figure 4 plots the Talagrand diagram (top), the sum of 108 Talagrand diagrams (middle), and the MST rank histograms (bottom) for 196 European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble forecasts of Northern Hemisphere 500-hPa height fields scattered through 2001. There are 51 ensemble members for forecast lead times of 0 days (first column), 1 day (second column), 5 days (third column), and 10 days (last column). The solid horizontal lines are the expected value for each bin based on the ensemble size, and the dashed lines are the expected standard deviation of each bin based on the ensemble size and the number of samples. Note that the standard deviation lines are not included in the summed Talagrand diagrams (middle row) because they are meaningless in this context. Based on the Talagrand diagrams on the top row the ensemble appears reliable, at least at the longer lead times. Simply summing the Talagrand diagrams from points distributed every 10° longitude on latitude rings of 30°, 45°, and 60° North yields the familiar ''nonflat'' patterns. These look far from flat, but because spatial correlations in the error fields are *expected*, the histograms cannot be interpreted as if based on independent events (each con-

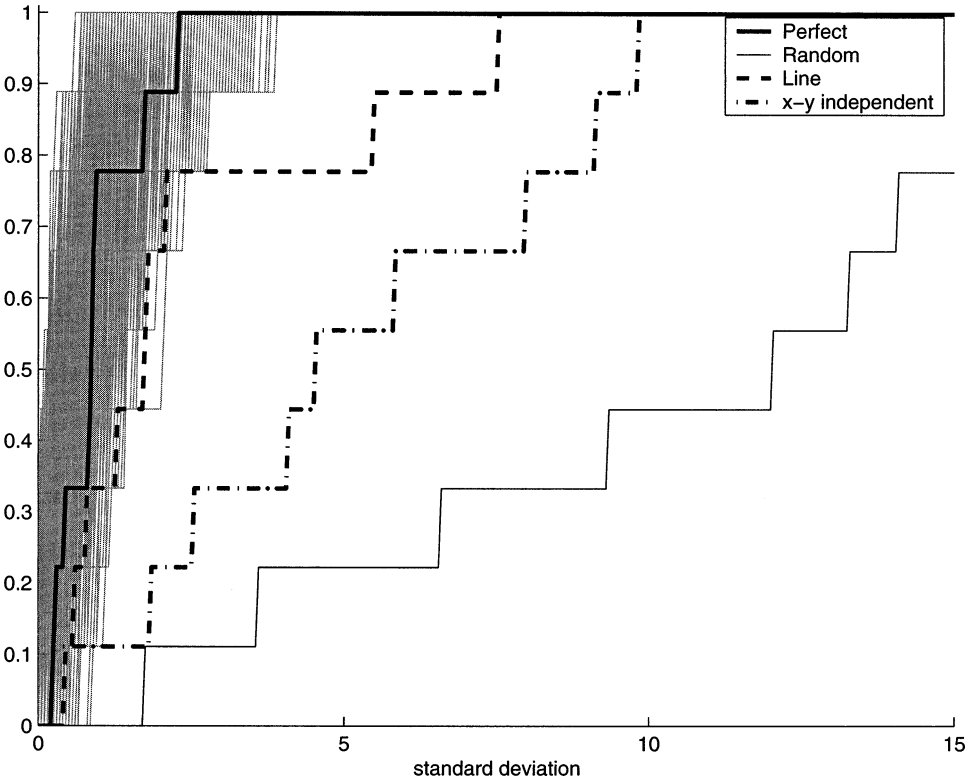

FIG. 3. Cumulative distribution function of departures from expected bin values normalized by expected std dev for the MST rank histograms in Fig. 2. The thick solid line is for verification drawn from the attractor, the thin solid line is for a randomly selected verification, the thick dashed line is for verification drawn from a line, and the thick dash–dot line is for x and y drawn independently from the correct distributions. The collection of thin gray lines are from 1000 realizations of random draws from a Gaussian distribution with mean zero and std dev 1.

tains 108 increments for every forecast). Thus the assumptions used above to define ''flat'' are violated, making it difficult to interpret this diagram when evaluating (or contrasting) ensemble forecast systems.

The MST rank histograms (bottom) are constructed in the full 69 173-dimensional space of the 500-hPa height field. It is apparent at 24 h that the ensembles are not reliable, and the reliability decreases with forecast lead. By 10 days one can see that verification systematically lies far from the ensemble members (the leftmost bins are overpopulated). Similar analysis has been performed on ensemble output from the Medium-Range Forecasting (MRF) model of the National Centers for Environmental Prediction (NCEP), and the National Center for Atmospheric Research (NCAR) Community Climate Model Version 3 (CCM3) with qualitatively similar results.

MST diagrams have application beyond ensemble assessment; they also provide a robust framework for the comparison of multidimensional distributions. Consider a long climate model run. A basis is defined based, say, on EOFs, and each January mean model state is projected onto each basis direction. Similarly, each observed January is projected onto the EOF same basis, thereby providing two high-dimensional distributions. The MST rank histogram approach can be used to determine the level at which *observed* January mean states are consistent with the model distribution as the number of EOFs (i.e., the dimension of the space) is increased.

While we hope the MST will prove a useful tool, we note that its strength in theory will prove some limitations in practice. Only in cases where both the model and the ensemble are perfect[6] can one expect to produce perfectly reliable ensemble forecasts (e.g., 30% probability events are observed to verify with a relative frequency of 0.3).

By design, the MST will detect instances where the verification is not drawn from the same distribution as the ensemble (hence the requirement of a perfect ensemble) or instances where the ensemble is not evolved under a perfect model. Current operational ensemble formation schemes were not designed to meet this goal and no model of a physical dynamic system is perfect.

While we can aim to improve our ensembles, and work using ensemble-based approaches to data assim-

[6] A perfect ensemble is one that is drawn from the same conditional distribution as the verification (Smith 1995).

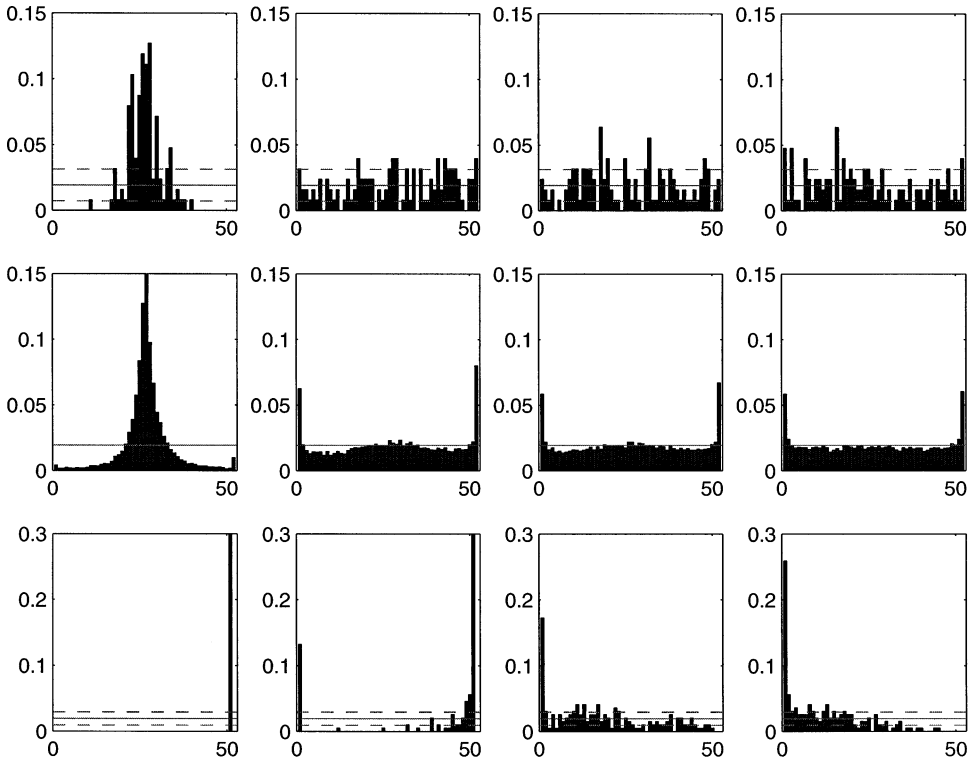

FIG. 4. (top) Talagrand diagrams, (middle) summed Talagrand diagrams from 108 locations, and (bottom) MST rank histograms for the ECMWF EPS for Northern Hemisphere 500-hPa height fields. Note that for the given sample size, one cannot say that the ensemble forecasts are unreliable when using the Talagrand diagram. Ensemble forecasts from the summed Talagrand diagrams have overpopulated end bins, but because of spatial correlations in the forecast error fields the histograms cannot be interpreted as if they were based on independent events. The MST rank histograms contain as many realizations as the Talagrand diagrams, yet the unreliability of the ensemble forecasts is evident after only 24 h. For convenience, the *y* axes for the MST rank histograms have been limited to 0.3 even though bin values exceed that level for the analyses and the 24-h forecasts.

ilation already show progress in the direction (Hansen 2002), we cannot require perfect MST diagrams from imperfect models any more than we can require perfect point forecasts from uncertain initial conditions. To do so would abuse the MST in a way not dissimilar to the way root-mean-square error statistics have been misused to overtune models in the past.

That said, the MST diagrams can provide a useful tool with wide application in the imperfect model context, both in terms of evaluating ensemble forecasts and in comparing distributions in high-dimension spaces. It should allow the improvement of probability forecasts. Fortunately, increasing the utility of a probability forecast does not require perfection, but merely that it prove more useful in application.

Acknowledgments. The authors thank K. Judd, A. Mees, and R. Smith for valuable conversations about ensemble assessment, and thank the Newton Institute at Cambridge University. The comments of two anonymous reviewers greatly enhanced the manuscript. This work was supported by ONR Predictability DRI Grant N00014-99-1-0056 and by ONR YIP N00014-02-1-0473.

REFERENCES

Anderson, J. L., 1996: Selection of initial conditions for ensemble forecasts in a simple perfect model framework. *J. Atmos. Sci.,* **53,** 22–36.

Efron, B., and R. J. Tibshirani, 1993: *An Introduction to the Bootstrap. Monographs on Statistices and Applied Probability,* Vol. 57, Chapman and Hall, 436 pp.

Epstein, E. S., 1969: Stochastic dynamic prediction. *Tellus,* **21,** 739–759.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.,* **129,** 550–560.

Hansen, J., 2002: Accounting for model error in ensemble-based state estimation and forecasting. *Mon. Wea. Rev.,* **130,** 2373–2391.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.,* **124,** 1225–1242.

Ikeda, K., 1979: Multiple-valued stationary state and its instability of the transmitted light by a ring cavity system. *Opt. Commun.,* **30,** 257–263.

Kruskal, J. B., 1956: On the shorted spanning subtree of a graph and

the traveling salesman problem. *Proc. Amer. Math. Soc.*, **7,** 48–50.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102,** 409–418.

Lorenz, E. N., 1965: Study of the predictability of a 28-variable atmospheric model. *Tellus*, **17,** 321–333.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF Ensemble Prediction System: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122,** 73–119.

Murphy, A. H., and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115,** 1330–1338.

Oseledec, V. I., 1968: A multiplicative ergodic theorem: Lyapunov characteristic numbers for dynamical systems. *Trans. Mosc. Math. Soc.*, **19,** 197–231.

Palmer, T. N., 2000: Predicting uncertainty in forecasts of weather and climate. *Rep. Prog. Physics*, **63** (2), 71–116.

Prim, R. C., 1957: Shortest connection networks and some generalizations. *Bell Syst. Tech. J.*, **36,** 1389–1401.

Smith, L. A., 1995: Accountability and error in nonlinear forecasting. *Proc. ECMWF Workshop on Predictability*, Shinfield Park, Reading, United Kingdom, ECMWF, 351–368.

——, 2000: Disentangling uncertainty and error: On the predictability of nonlinear systems. *Nonlinear Dynamics and Statistics*, A. I. Mees, Ed., Birkhauser, 31–64.

——, 2003: Predictability past predictability present. *Proc. ECMWF Seminar on Predictability*, Shinfield Park, Reading, United Kingdom, ECMWF, 219–242.

——, C. Ziehmann, and K. Fraedrich, 1999: Uncertainty dynamics and predictability in nonlinear systems. *Quart. J. Roy. Meteor. Soc.*, **125,** 2855–2886.

Talagrand, O., R. Vautard, and B. Strauss, 1999: Evaluation of probabilistic prediction systems. *Proc. ECMWF Workshop on Predictability*, Shinfield Park, Reading, United Kingdom, ECMWF, 1–25.

Thompson, P. D., 1957: Uncertainty of initial state as a factor in the predictability of large-scale atmospheric flow patterns. *Tellus*, **9,** 275–295.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74,** 2317–2330.

Van den Dool, H. M., 1994: Searching for analogues, how long must we wait? *Tellus*, **46A,** 314–324.

von Mises, R., 1981: *Probability, Statistics and Truth.* Dover, 244 pp.

Ziehmann, C., L. A. Smith, and J. Kurths, 2000: Localized Lyapunov exponents and the prediction of predictability. *Phys. Lett.*, **271A,** 237–251.