Extending a Metric Developed in Meteorology to Hydrology

D.M. Sedorovich[1]
[1]Department of Agricultural and Biological Engineering, The Pennsylvania State University, University Park, PA 16802, USA
dms521@psu.edu

Abstract

All model analyses must include some measure of prediction accuracy regardless of the application. Many metrics exist, with no one best universally applicable measure. This paper presents a modification of the forecast quality index (FQI), a metric developed in meteorology. This metric, FQI_m, was applied to a simple case study and found to be applicable to hydrological systems. It includes both an amplitude- and distance-based component, although it is more advantageous to use the components separately as the hybrid measure is more impacted by the distance-based component. The metric is useful for measuring prediction accuracy, calibrating a model using ensembles, and identifying the trade-off between multiple model objectives.

INTRODUCTION

Computer models are an inexpensive and relatively quick method of analyzing different scenarios or predicting possible outcomes. Despite this benefit, there are many issues associated with modeling natural systems. How do you decide if a model accurately simulates your system? Which metric is most appropriate for your application? At what value of a metric is your model "good enough"? Can metrics used in one field be applied to another?

Although there is no one correct answer to these questions, various disciplines have approached these issues differently. This paper attempts to provide insight into the last question by investigating whether a metric developed for use in meteorology, the forecast quality index (FQI), can be applied to hydrology. Related to this goal, this paper will also address the first three questions by analyzing what FQI value corresponds to the minimal residual errors and by determining how appropriate the FQI is to hydrological models. To accomplish these goals, this paper will first discuss the role of metrics in modeling and some commonly used methods in meteorology. Then, it will describe the FQI, how hydrology differs from meteorology, and modifications necessary for hydrological applications. Finally, a simple case study will be used to illustrate the usefulness of the modified metric to hydrology.

VALIDATION METHODS IN METEOROLOGY

Ensemble Prediction Systems (EPS)

Ensemble prediction systems are stochastic simulations with members consisting of either different model structures or different input parameters. Unlike deterministic simulations that result in a single output, EPS result in a distribution of outputs. This distribution can be used for three purposes. First, the distribution can be used to identify the probability of a given outcome occurring (e.g., the probability that stream flow will exceed a threshold). Second, the distribution provides an indication of the variability resulting from different parameter values. From a different perspective, this second purpose can be restated as determining how errors in the input values propagate errors in the output. Finally, the distribution can be used in model calibration to identify the parameter set that minimizes the errors between the simulated and observed value based on a selected objective function. There are a variety of metrics that can be used, with the selection dependent on the goal of the calibration.

Overview of Metrics

Typical metrics familiar to hydrologists that are also used in meteorology include the root mean square error (RMSE), mean absolute error (MAE), Nash-Sutcliffe efficiency (NSE), and the correlation coefficient (R^2). In addition, there are certain metrics that are more commonly used in meteorology as compared to other fields such as Talagrand diagrams (Talagrand et al., 1997), reliability diagrams (Ebert et al., 2005), Brier skill score (Ebert et al., 2005), and the Wilson method (Wilson et al., 1999). As in other fields, none of these measures is a universal best metric to determine the accuracy of all models. Each of these measures is either a distance- or amplitude-based measure, whereas most validation attempts require a measure of both.

Forecast Quality Index (FQI)

Venugopal et al. (2005) proposed a hybrid measure for validation of meteorological models and tested it by comparing the percentage of observed and predicted precipitation covered area. This measure, the forecast quality index (FQI), consisted of both amplitude- and distance-based components.

$$FQI(R_1, R_2) = \frac{\dfrac{PHD_k(R_1, R_2)}{Mean\left[PHD_k(R_1, Surrogates\ R_1)\right]}}{\dfrac{2\mu_{R_1}\mu_{R_2}}{\mu_1^2 + \mu_2^2}\ \dfrac{2\sigma_{R_1}\sigma_{R_2}}{\sigma_1^2 + \sigma_2^2}} \tag{1}$$

The numerator represents the distance-based component and is calculated using a normalized Partial Hausdorff distance (PHD). The PHD measures the similarity between *R₁* and *R₂* by measuring the distance between all points in *R₁* and all points in *R₂*.

$$H(A, B) = \max\left(\min\|a - b\|\right) \tag{2}$$

The norm ‖a-b‖ can be determined through various methods including the absolute difference (‖a-b‖ = abs(‖a‖-‖b‖)), Euclidean distance ($\|a\text{-}b\|^2 = \|a\|^2 + \|b\|^2$), or “taxi-cab” distance (‖a-b‖ = ‖a‖+‖b‖), among others. The PHD was normalized by the PHD of R_1 and its surrogates to account for the wide variation possible in the percentage of nonzero pixels over time or between ensemble members. The surrogates were generated by using the iterative amplitude-adjusted Fourier transform (IAAFT) algorithm of Schreiber and Schmitz (1996) which preserves both the correlation structure and the probability density function or *R₁* (for a more in-depth discussion of calculating surrogates, see Kantz and Schreiber (1997) and Schreiber and Schmitz (1996)).

The denominator represents the amplitude-based component and is calculated using a modified universal image quality index (UIQI). The UIQI (Wang and Bovik, 2002) consisted of components for the correlation, brightness (bias), and distortion (variability).

$$UIQI(R_1, R_2) = \frac{\sigma_{R_1,R_2}}{\sigma_{R_1}\sigma_{R_2}} \frac{2\mu_{R_1}\mu_{R_2}}{\mu_1^2 + \mu_2^2} \frac{2\sigma_{R_1}\sigma_{R_2}}{\sigma_1^2 + \sigma_2^2} \qquad (3)$$

where σ_{R_1,R_2} is the covariance between R_1 and R_2, σ_{R_1} and σ_{R_2} are the standard deviations of the two fields, and μ_{R_1} and μ_{R_2} are the means of the two fields. The correlation term is accounted for in the PHD and, as a result, Venugopal et al. (2005) only use the brightness and distortion terms in the FQI.

HYDROLOGICAL APPLICATION

Modified FQI (FQI$_m$)

Modifications were made to make equation (1) more suitable to hydrology and the specific application being addressed in this paper.

$$FQI_m(R_1, R_2) = \frac{HD(R_1, R_2)}{\frac{2\mu_{R_1}\mu_{R_2}}{\mu_1^2 + \mu_2^2} \frac{2\sigma_{R_1}\sigma_{R_2}}{\sigma_1^2 + \sigma_2^2}} \qquad (4)$$

where FQI_m is the modified FQI to distinguish it from the published FQI, *HD* is the classical Hausdorff distance, and the remaining variables are as defined previously.

By comparing equation (1) and (4), it is obvious that the main modifications are in the numerator. First, the 100th percentile (k = 100) was used instead of the 75th (k = 75). By using the 100th percentile, the numerator of equation (1) reduced to the classical definition of the

Hausdorff distance (HD) as opposed to the partial distance. The classical definition of the HD is more sensitive to outliers than the PHD. However, correctly simulating the extreme values in hydrology (e.g., maximum annual stream flow) was deemed important enough to justify the increased sensitivity to outliers. Second, the numerator was not normalized because of the different application of FQI_m used in this project. The application in Venugopal et al. (2005) compared multiple images throughout time, resulting in a time variance issue (i.e., the variation throughout time of the percent of precipitation covered area). The application for this paper did not have the time variance issue and thus did not require normalization of the HD. Finally, Venugopal et al. (2005) used the "taxi-cab" distance as the norm in equation (2) whereas equation (4) used the Euclidean distance, with the daily stream flow and the day as the "coordinates."

The numerator of FQI_m ranges from $0 - \infty$, with the best value being 0; the denominator of FQI_m ranges from $0 - 1$, with the best value being 1. As a result, FQI_m ranges from $0 - \infty$, with the best value being 0.

Methodology

The FQI_m was tested using the Leaf River data and the HyMod01 model. The Leaf River data consisted of observed precipitation, potential evapotranspiration, and stream flow for multiple years. The HyMod01 model consisted of a probability distribution model for soil moisture accounting, a Nash cascade to route quick flow, and an infinite linear tank to route slow flow. For each step, the analyses and conclusions are based on a 365-day simulation. However, only a 50-day section of the time series is shown in the results section in order to highlight the differences between ensemble members. The methodology followed is applicable to time series of any length.

Three steps were followed to determine the suitability and usefulness of the FQI_m measure to hydrologic data:

(a) Testing of FQI_m on perturbed data

(b) Calibration of the HyMod01 model using FQI_m

(c) Objective functions (OF) comparison: FQI_m, $FQI_{m,N}$, $FQI_{m,D}$, RMSE, and NSE

Perturbed data The goal of this step was to confirm that the values of FQI_m changed as expected. In other words, if the amplitude of an ensemble member exactly matched that of the observed data for the entire time simulated, the denominator ($FQI_{m,D}$) was expected to be one; if the phases of an ensemble member and the observed data matched for the entire time simulated (i.e., there was no lag in the simulated data), the numerator ($FQI_{m,N}$) was expected to be zero. The expected values were tested through three scenarios:

(a) Amplitude shifted – the amplitude of the entire hydrograph was shifted by a random amount (positive or negative), leaving the phase the same; expected value: $FQI_{m,N} = 0$, $FQI_{m,D} =$ varies, with members closer to the observed data closer to one.

(b) Phase shifted – the entire hydrograph was shifted forward or backward, leaving the amplitude the same; expected value: $FQI_{m,D} = 1$, $FQI_{m,N} =$ varies, with members with less of a lag closer to 0.

(c) Amplitude and phase shifted – the amplitude and phase were shifted by a random amount; ensemble members closer in shape to the observed data were expected to have $FQI_{m,N}$ approaching zero, $FQI_{m,D}$ approaching one, and FQI_m approaching zero.

Calibration The goal of this step was to determine how FQI_m could be applied for calibrating models. Five parameters were chosen for calibration, two for the soil moisture accounting (H_{uz}, and B) and three for the flow routing (K_q, K_s, and *alpha*). A 10 000 member ensemble was generated consisting of different input parameter sets rather than different model structures. Typically, ensemble members are generated using singular or breeding vector approaches (Lewis, 2004) instead of non-random sampling of the probability density function (pdf). However, for this application, a Monte Carlo sampling of the five parameters based on a uniform random distribution was used to produce the ensemble. Each parameter set (i.e., ensemble member) was then used as inputs to HyMod01, generating a distribution of predicted stream flow data. The best ensemble members were chosen based on the best value of $FQI_{m,N}$, $FQI_{m,D}$, and FQI_m, and compared to those chosen based on the RMSE and NSE.

OF The goal of this step was to determine how FQI_m differed from other commonly used OF. The OF used were FQI_m, FQI_{m-N}, FQI_{m-D}, RMSE, and NSE. The same procedure used to generate an ensemble for the calibration step was followed and each objective function calculated. The five objective functions were then plotted against each other to show the trade-off curve.

Results

Perturbed data Figure 1 shows the three cases of the perturbed data with the error statistics found in Table 1. For the perturbed amplitude, ensemble member two (Amp 2) was better than member one for all statistics. For the perturbed phase, member one (Phase 1) was better for all statistics. For the perturbed amplitude and phase, ensemble member one (PA 1) was better than member two for all statistics except $FQI_{m,D}$.

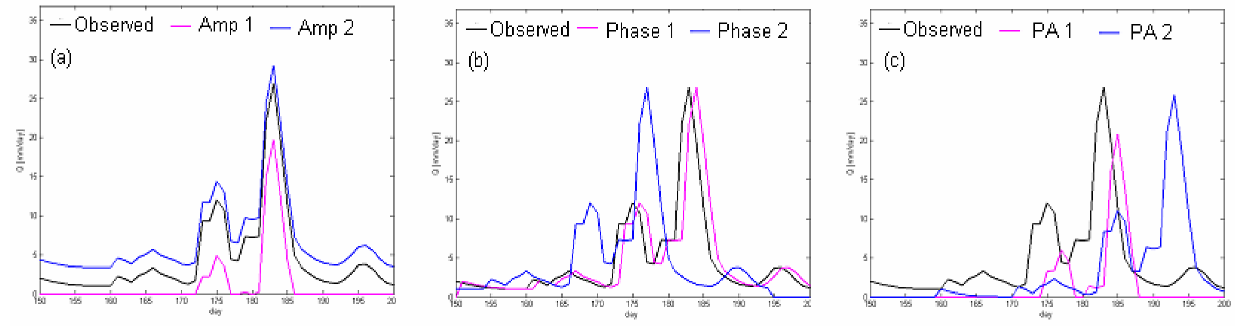

Figure 1. Observed data and two ensemble members for the Leaf River from day 150 - 200. (a) Perturbed amplitude (b) Perturbed phase (c) Perturbed amplitude and phase.

Table 1. Comparison of error metrics for six ensemble members.

	Best Value	Amp 1	Amp 2	Phase 1	Phase 2	PA 1	PA 2
FQI_m	0	10.08	2.57	2.01	6.01	8.59	10.59
$FQI_{m,N}$	0	5.00	2.36	2.01	6.01	5.05	10.05
$FQI_{m,D}$	1	0.50	0.92	1.00	1.00	0.59	0.95
NSE	0	0.28	0.57	0.41	-0.35	-0.11	-0.36
RMSE	1	3.92	2.36	3.24	7.38	6.06	7.48

Calibration Figure 2a shows the observed stream flow and distribution of the 10 000 ensemble members. The ensemble had a wide distribution around the observed data. Figure 2b shows the observed stream flow and the five best of the 10 000 ensemble members based on different metrics. The best members were observed to vary based on the different metric used.

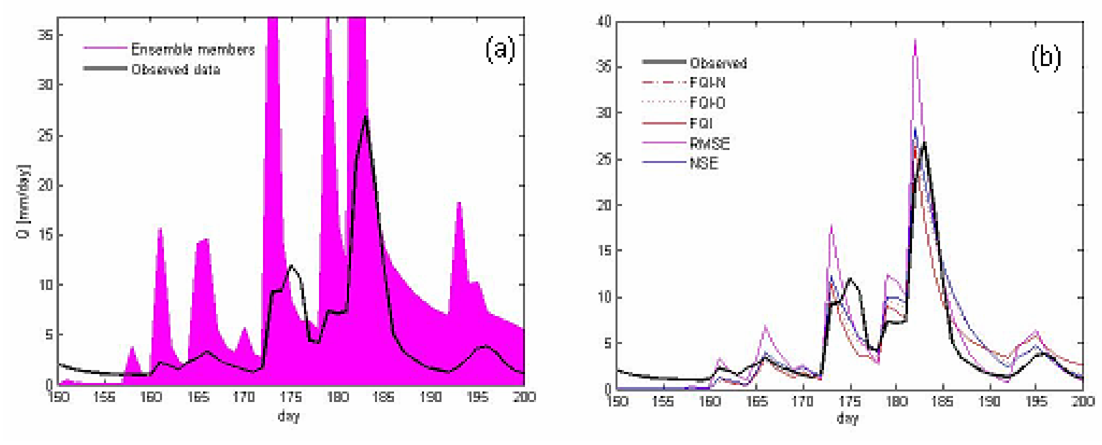

Figure 2. 10 000 member ensemble results. (a) Stream flow distribution of ensemble members and observed stream flow. (b) Observed and best members based on $FQI_{m,N}$, $FQI_{m,D}$, FQI_m, RMSE, and NSE.

OF Figure 3 shows the graphs of OF vs. OF for FQI_m, $FQI_{m\text{-}N}$, $FQI_{m\text{-}D}$, RMSE, and NSE, with the best value of each found in Table 1. The graph for $FQI_{m\text{-}N}$ vs. $FQI_{m\text{-}D}$ is not shown because the same information can be found in the graphs with the combined metric, FQI_m.

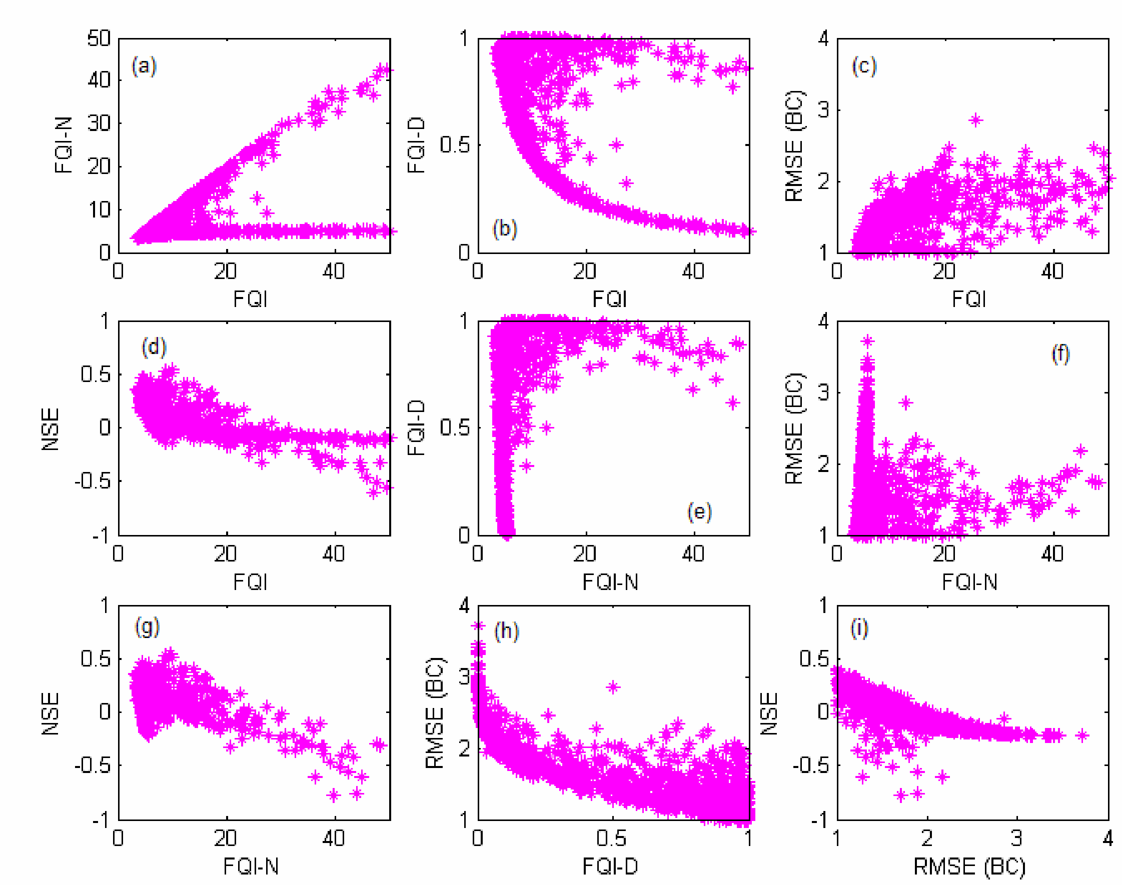

Figure 3. Objective function graphs for FQI_m, $FQI_{m\text{-}N}$, $FQI_{m\text{-}D}$, RMSE, and NSE.

Discussion

When the amplitude of an ensemble member is closer to the observed amplitude, $FQI_{m\text{-}D}$ approaches one, as shown by the results for Amp1 and Amp2. When the hydrograph does not include any lag (i.e., is not shifted) $FQI_{m\text{-}N}$ approaches zero, as shown by the results for Phase1 and Phase2. For these two ensemble members, the amplitude was not perturbed and $FQI_{m\text{-}D}$ for both members is one as expected. One drawback is that the $FQI_{m\text{-}N}$ does not have an upper bound and, as a result, ensemble members must be compared to each other to determine which is better. The metric thus provides a relative estimate of the magnitude of the difference between each ensemble member and the observed data. When analyzing these four ensemble members, it is obvious that the metric does not indicate the direction of the difference. From members PA1 and PA2, it is obvious that FQI_m is slightly biased toward the distance component (FQI_{m-N}). Member PA1 was chosen as the better member based on FQI_m. The amplitude of PA1 differs from the observed more than PA2 although PA1 has less of a lag than PA2. This is likely due to the different range of FQI_{m-N} ($0 - \infty$) compared to FQI_{m-D} ($0 - 1$), resulting in the numerator impacting the overall metric more than the denominator. Based on these results, the metric was confirmed to behave as expected for hydrologic data. Ideally the bias toward the distance component would be removed or at least reduced; the metric can be improved by normalizing FQI_{m-N} so that it does not have an undue contribution to FQI_m. The metric would also be improved if it indicated whether the amplitude was under- or over-predicted and whether the phase shift was forward or backward.

The calibration results show a useful intersection between EPS and FQI_m. Figure 2a can be used for either of the first two purposes discussed in the EPS section. In addition, combining the distribution with FQI_m can be used for the third purpose. Choosing the ensemble member with the best value of FQI_m can identify the parameter set that minimizes the residuals between observed and predicted stream flow. Using different metrics (e.g., FQI_{m-N}, FQI_{m-D}, RMSE, NSE) would favor a different aspect of the hydrograph (e.g., FQI_{m-N} minimizes lag), resulting in a different optimal parameter set.

The OF results can be used to find the trade-off between different objectives. For example, Figure 3c shows the variation in FQI_m and RMSE over ensemble members. The best parameter sets for these two objectives are represented by points in the lower left corner of the graph. Optimizing for more than two objectives is possible but would require different visualization techniques than are used here and is out of the scope of this paper.

Four general conclusions and avenues of further research on FQI_m can be made from the specific results described above. First, in its current form, the bias of the FQI_m toward the lag indicates that the metric proposed by Venugopal et al. (2005) does not provide any new information that could not be obtained separately from the PHD and UIQI. It would be more advantageous to use the two components individually rather than combining them into one hybrid metric. Second, it is easy to visually confirm that a simulated hydrograph has the correct shape (i.e., no lag) but it is one of the most difficult aspects to mathematically quantify. Metrics that quantify this aspect are ideal objectives to use for validation and calibration of models. Based on this, the FQI_{m-N} is determined to be the more important of the two components of FQI_m. Following from the first conclusion, it is recommended to use FQI_{m-N} in combination with another metric that is partial to the amplitude (e.g., FQI_{m-D}, NSE). Third, normalizing FQI_{m-N} and modifying the FQI_m to reduce the bias would likely improve the usefulness of the metric. Finally, it would be interesting to compare FQI_m results from an application with real data to one with binary data. The original FQI was applied to binary data whereas this application used real data. Hydrologic data can easily be converted to binary format, for example, by defining a stream flow threshold with flows above the threshold designated as an event occurrence.

CONCLUSION

A metric developed in meteorology has been modified and applied to hydrological models. The FQI_m measures how accurately a model simulates both the amplitude and the phase of the observed hydrograph. However, the most advantageous application of the FQI_m is by analyzing its components separately rather than as a hybrid measure. Specifically, the distance-based component, FQI_{m-N}, was found to be useful for hydrological applications because of its ability to mathematically quantify a visual aspect of stream flow, namely whether the shape of the hydrograph is accurately simulated. It is recommended to combine the FQI_{m-N} with a commonly accepted metric such as the NSE to provide a measure of how accurately the model predicts both amplitude (NSE) as well as phase (FQI_{m-N}). The metric can be further improved by reducing the bias to distance and by normalizing the distance-based component.

REFERENCES

Ebert, B., Brown, B., Wilson, L., Nurmi, P., Brooks, H., Casati, B., Stephenson, D., Gober, M., Ghelli, A., Damrath, U., Wilson, C., Baldwin, M., Brill, K., Joliffe, I., & Atger, F. (2005). Forecast verification: Issues, Methods, and FAQs. WWRP/WPGNE Joint Working Group on Verification. Available at http://www. bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html.

Huttenlocher, D.P., Klanderman, G.A., & Rucklidge, W.J. (1993) Comparing images using the Hausdorff distance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(9):951 – 956.

Kantz, H. & Schreiber, T. (1997) Nonlinear time series analysis. Cambridge Univ. Press, New York.

Lewis, J. (2005) Roots of ensemble forecasting. *Mon. Wea. Rev.*, 133(7):1865 – 1885.

Schrieber, T. & Schmitz, A. (1996) Improved surrogate data for nonlinearity tests. *Phys. Rev. Lett.*, 77:635 – 638.

Talagrand, O., R. Vautard, and B. Strauss. 1997. Evaluation of probabilistic prediction systems. *Proceedings:ECMWF Workshop on Predictability.*

Taylor, K.E. 2001. Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.*, 106(D7):7183 – 7192.

Venugopal, V., Basu, S. & Foufoula-Georgiou, E. (2005) A new metric for comparing the precipitation patterns with an application to ensemble foreceasts. *J. Geophys. Res.*, 110, D08111, doi:10.1029/2004JD005395.

Wang, Z. & Bovik, A.C. (2002) A universal image quality index. *IEEE Signal Process. Lett.*, 9(3):81 – 84.

Wilson, L.J., W.R. Burrows, and A. Lanzinger. 1999. A strategy for verification of weather element forecasts from an ensemble prediction system. *Mon. Wea. Rev.,* 127:956 – 970.