14A.7 VERIFICATION OF DETERMINISTIC TROPICAL CYCLONE INTENSITY FORECASTS: MOVING BEYOND MEAN ABSOLUTE ERROR

Jonathan R. Moskaitis*
Massachusetts Institute of Technology, Cambridge, Massachusetts

1. INTRODUCTION

Deterministic forecasts of tropical cyclone (TC) intensity are typically verified by calculating a *summary accuracy measure*, such as mean absolute error or mean squared error. These summary measures quantify, with a scalar value, the quality of the relationship between a set of forecasts and the corresponding set of observations (i.e. "forecast quality"). While a useful simplification, representation of forecast quality with a summary accuracy measure does not respect the complexity of the relationship between forecasts and observations. To obtain a complete representation of forecast quality, the joint probability distribution of forecasts and observations must be estimated. Here, such joint distributions are estimated for operational forecasts of TC intensity (produced by the NHC, Decay-SHIPS, SHIFOR and GFDL), using 5 recent years of Atlantic basin forecasts and the corresponding best track observations. Analysis of the joint distributions shows that for all forecast systems/agencies, intensity predictions tend to asymptote with lead time toward the central tendency of the climatological intensity distribution. It is argued that this behavior is a response to the demand that operational TC intensity forecast systems minimize the mean absolute error of their predictions.

The organization of this paper is as follows. First, mean absolute error verification and the results of its application to operational TC intensity forecasts are reviewed in Sec. 2. Subsequently in Sec. 3, estimates of the joint probability distributions of operational TC intensity forecasts and the corresponding observations are presented. Sec. 4 links the first-order nature of the joint distributions to the demand of mean absolute error minimization. Finally, a brief summary of the work described herein and conclusions are contained in Sec. 5.

2. MEAN ABSOLUTE ERRORS

Consider a verification data sample, denoted $\{(f_k, x_k); k = 1, \ldots, N\}$, consisting of N forecasts, f, and the corresponding observations, x. In this context, mean absolute error is defined as

$$\mathrm{MAE} = \frac{1}{N}\sum_{k=1}^{N} |f_k - x_k|. \tag{1}$$

Fig. 1 shows the MAE for four operational TC intensity forecast systems, at eight lead times ranging from 0 to 120 hours (32 total verification data samples). The forecast systems/agencies represented are the National Hurricane Center forecasters (abbreviated "OFCL", the official forecast), the 5-day Statistical Hurricane Intensity Forecast model (SHF5), the Decay-Statistical Hurricane Intensity Prediction Scheme (DSHP), and the GFDL/URI coupled hurricane-ocean model (GFDL). For each forecast system at each lead time, the verification data sample is composed of intensity predictions for Atlantic basin TCs (from the 2001 through 2005 seasons) and the corresponding "observations" from the best track analysis. To facilitate fair comparison amongst the forecast systems, only cases where all four forecasts exist and can be compared against an existing best track observation are included in the verification data samples[1].

Fig. 1 is the standard format for displaying/comparing the performance of TC intensity forecast systems (e.g. Franklin cited 2007; Knaff et al. 2003; DeMaria et al. 2005; Bender et al. 2007, for OFCL, SHF5, DSHP, and GFDL forecasts, respectively). To first order, TC intensity forecast systems are driven to produce the lowest possible line on this plot, as the forecast system with the lowest MAE is considered the "best". Thus, efforts to improve of the quality of predictions from an intensity forecast system proceed with the goal of MAE minimization,

* *Corresponding author address*: Jonathan R. Moskaitis, Rm. 54-1611, MIT, 77 Massachusetts Ave., Cambridge, MA 02139; *e-mail: jonmosk@mit.edu.*

[1] Further details concerning the verification data samples (and this work in general) can be found in the full manuscript entitled "A case study of deterministic forecast verification: Tropical cyclone intensity", available at web.mit.edu/jonmosk/www.

without explicit regard to other attributes of forecast quality. To understand the consequences for overall forecast quality of this MAE minimization approach to intensity forecast system development, the joint probability distributions of operational TC intensity forecasts and the corresponding observations must be analyzed.

3. JOINT PROBABILITY DISTRIBUTIONS

To fully appreciate the complexity of a verification data sample, as embodied in a joint distribution of forecasts and observations, it is useful to work backwards from the familiar MAE expression of Eq. 1. The MAE takes the average of a collection of absolute errors, $\{|e_k| = |f_k - x_k|; k = 1, \ldots, N\}$, resulting in a scalar value. Instead of taking this average, one could imagine plotting the relative frequency distribution of the $|e_k|$ to investigate how often each magnitude absolute error occurs in the verification data sample. Better yet, the relative frequency distribution of the e_k could be plotted, to distinguish between errors of opposite sign, but the same magnitude. However, this still leaves considerable ambiguity about the relationship between the forecasts and observations. In terms of error, $(f_k, x_k) = (30\,\mathrm{kt}, 50\,\mathrm{kt})$ is indistinguishable from $(f_k, x_k) = (130\,\mathrm{kt}, 150\,\mathrm{kt})$, for example. To eliminate this sort of ambiguity, one must consider the joint relative frequency distribution of forecasts and observations, rather than a relative frequency distribution of errors. This joint relative frequency distribution distribution serves as an easily-obtainable estimate of the joint probability distribution, $p(f, x)$, and describes all the time-independent information that a verification data sample has to offer about the forecasts, the observations, and their relationship. It is the fundamental instrument of the "distributions-oriented" approach to verification (Murphy and Winkler 1987).

Simply plotting the joint distribution of forecasts and observations can lead to useful insights about forecast quality. In Fig. 2, the joint distribution of OFCL forecasts and best track observations is plotted at each of four different lead times. In each panel, dots are drawn for all (f, x) with non-zero relative frequency in the corresponding verification data sample, with the colors representing the magnitude of the relative frequency according to the nonlinear scale detailed below Fig. 2. Note that the joint distribution is discrete, such that (f, x) is indicative of certain ranges of f and x; for example $(f, x) = (70, 85)$ represents $67.5 \leq f < 72.5$ and $83.5 \leq x < 88.5$. Like Fig. 2, Figs. 3–5 show the joint distributions for the SHF5, DSHP, and GFDL, respectively. All such joint distributions are estimated based on the same verification data samples used to calculate the MAEs shown in Fig. 1.

If a sample of deterministic forecasts were perfect, its joint distribution with the corresponding sample of observations would show dots only along the diagonal, $f = x$. Figs. 2–5 demonstrate that this is not the case for any of the TC intensity forecast samples, even at the 0 h lead time (because operational analyses of intensity do not necessarily match the best track values). At the 36 h lead time, all four forecast samples show a widening of the joint distribution about the (diagonal) major axis, indicating a growing proportion of large forecast errors. By the 72 h lead time, the joint distributions have widened further about their major axes and the major axes have rotated into a more vertical orientation. Finally, at the 120 h lead time, the rotation of the major axes into the vertical is more evident, and if anything, the joint distributions have gathered in towards their major axes rather than spreading out further.

It is important to reiterate that the evolution of the joint distributions in lead time described above is common to all four forecast systems. Such evolution signifies the same type of deficiency in forecast quality amongst the forecast systems: conditional bias[2] that grows in magnitude with lead time. The clearest example of conditional bias is in the SHF5 forecasts at the 120 h lead time, as shown in Fig. 3d. For high intensity observations the corresponding forecasts are generally too low, and for low intensity observations the corresponding forecasts are generally too high. This pattern of conditional bias is present for all forecast systems for all positive lead times shown in Figs. 2–5, and is ultimately indicative of the influence of MAE minimization, as will be described subsequently.

4. DIAGNOSING THE INFLUENCE OF MAE MINIMIZATION

The conditional bias inferred from the joint distributions in Figs. 2–5 is a direct result of the changing nature of the marginal distributions of forecasts with lead time. Ideally, a marginal distribution of fore-

[2] Conditioning on the observation, specifically. This is called type II conditional bias.

casts,

$$s(f) = \sum_x p(f, x),$$

would match the corresponding marginal distribution of observations,

$$t(x) = \sum_f p(f, x),$$

implying that each intensity value is forecasted as often as it is observed. In Fig. 6, marginal distributions of operational TC intensity forecasts (dashed lines) are superimposed upon the marginal distribution of observations (solid line), each panel for a different lead time. For display purposes, continuous approximations to the discrete marginal distributions are shown. One can see that as lead time advances, the marginal distributions of forecasts sharpen relative to the marginal distribution of observations. This is because too many forecasts of moderate intensity TCs are made while too few forecasts of strong and weak TCs are made (relative to observed intensity occurrences), a phenomenon that becomes more pronounced with lead time. The sharpening of the marginal distributions of forecasts is manifested in the joint distributions of Figs. 2–5 as a rotation of the major axis from the diagonal into a more vertical orientation and an attendant contraction of the distribution about the more vertical major axis. Such an evolution of the features of the joint distribution is necessary to accommodate the sharpening of the marginal distribution of forecasts.

The sharpening of the marginal distributions of forecasts as lead time advances is a ultimately a consequence of MAE minimization. This can be understood by viewing deterministic TC intensity prediction within the context of probabilistic TC intensity prediction, the theoretically-correct approach. Although there is uncertainty in the intensity of a TC at the 0 h lead time, it can be confidently asserted that the true intensity is drawn from a probability distribution that is much sharper than the climatological intensity distribution (approximated by the marginal distribution of observations in Fig. 6d). For example, one might suppose that the true intensity is drawn from a Gaussian with a small (5 kt, perhaps) standard deviation, centered on the operationally-analyzed intensity value. However, as forecast lead time increases, uncertainty in the TC intensity increases, eventually to the point where it cannot be claimed that the probability of realizing any particular intensity value is different from the climatological probability. Deterministic TC intensity forecast systems have learned, through either implicit or explicit means, to respond to this inherently probabilistic prediction context in a manner that minimizes the expected absolute error of their deterministic forecasts. The protocol is to always forecast the median of the probability distribution for intensity. At the 0 h lead time, this results in a deterministic forecast of the operationally-analyzed intensity (assuming a symmetric probability distribution about that value, as in the example above), and at very long lead time, this results in a forecast of the median of the climatological intensity distribution. Hence, an absolute error-minimizing deterministic forecast trajectory starts at the operationally-analyzed intensity value and asymptotes toward the median of the climatological intensity distribution as lead time increases. The sharpening of the marginal distribution of forecasts with lead time in Fig. 6 is a feature of a collection of the aforementioned type of deterministic forecast trajectories.

5. SUMMARY AND CONCLUSIONS

Here, deterministic TC intensity forecasts have been verified using a distributions-oriented approach, which is based on analysis of the joint probability distribution of forecasts and observations. Relative to summary accuracy measure verification, distributions-oriented verification may seem cumbersome[3], but it gives a much more complete picture of forecast quality. For operational TC intensity forecasts, analysis of the joint distributions revealed an increasing (type II) conditional bias with lead time to be the primary deficiency in forecast quality. The conditional bias was linked to a sharpening of the marginal distribution of forecasts with lead time, itself the ultimate result of the demand of mean absolute error minimization imposed on deterministic TC intensity forecast systems. It is important to note that this result is common to all four of the forecast systems/agencies evaluated here, representing statistical models, a dynamical model, and NHC forecasters. While differing substantially in the methodology of forecast production, these forecast system/agencies share the goal of minimizing the MAE of their intensity predictions.

In conclusion, it must be stressed that MAE minimization is not inherently a “bad” way to drive the

[3] For example, the 16 joint distributions shown in Figs. 2–5 account for only half the lead time/forecast system combinations for which the MAE is displayed in Fig. 1.

improvement of forecast systems, even though it condones conditionally biased forecasts. As with any other summary measure, MAE is simply limited in its ability to represent the rich relationship between forecasts and observations embodied in the joint probability distribution. Ultimately it would be best to evaluate the performance of forecast systems based on distributions-oriented verification techniques, but the complexity of probability distributions (joint or marginal) hinders straightforward objective comparison of those from competing forecast systems[4]. Utilizing multiple summary measures covering different attributes of forecast quality (accuracy, unconditional bias, conditional bias, etc.) is perhaps the next best option in forecast quality assessment. In verification of TC intensity forecasts, a summary measure of the *information content* attribute of forecast quality is of particular interest. This measure, the mutual information between the forecasts and observations (DelSole 2005), quantifies the average amount of information a forecast provides about the observation, relative to prior knowledge of sample climatology. Forecasts/observations of a dissipated TC can be included in data samples verified with mutual information, whereas such forecasts/observations are necessarily excluded in MAE verification (as $f = 50\,\mathrm{kt}$ minus $x = \mathrm{dissipated}$ is not meaningful, for example). Such mutual information verification of data samples including forecast/observation realizations involving TC dissipation is demonstrated in Moskaitis (2008).

Acknowledgement

The author wishes to thank Kerry Emanuel and James A. Hansen for their encouragement in this line of research, and acknowledge the support of ONR Grant N00014-05-1-0323.

References

Bender, M. A., I. Ginis, R. Tuleya, B. Thomas, and T. Marchok, 2007: The operational GFDL coupled hurricane-ocean prediction system and summary of its performance. *Mon. Wea. Rev.*, **135**, 3965–3989.

DelSole, T., 2005: Predictability and information theory. Part II: Imperfect forecasts. *J. Atmos. Sci.*, **62**, 3368–3381.

DeMaria, M., M. Mainelli, L. K. Shay, J. A. Knaff, and J. Kaplan, 2005: Further improvements to the Statistical Hurricane Intensity Prediction Scheme (SHIPS). *Wea. Forecasting*, **20**, 531–543.

Franklin, J., cited 2007: National Hurricane Center forecast verification. [Available online at http://www.nhc.noaa.gov/verification].

Knaff, J. A., M. DeMaria, C. R. Sampson, and J. M. Gross, 2003: Statistical, 5-day tropical cyclone intensity forecasts derived from climatology and persistence. *Wea. Forecasting*, **18**, 80–92.

Moskaitis, J. R., 2008: An information-theoretic approach to quantifying the uncertainty in operational tropical cyclone intensity predictions, with application to forecast verification. *Preprints, 88th Annual Meeting: Tropical Meteorology Special Symposium*, New Orleans, LA, Amer. Meteor. Soc., J4.3.

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

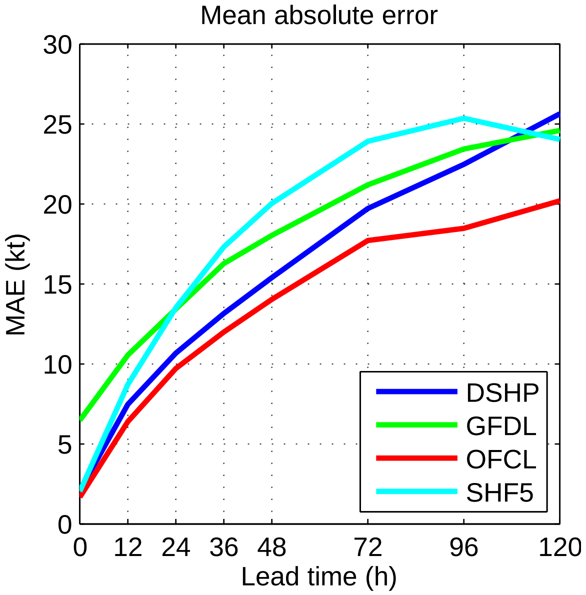

Figure 1: Mean absolute error, as a function of lead time, for the OFCL (red), SHF5 (light blue), DSHP (dark blue) and GFDL (green) intensity forecast systems.

[4] Even subjective comparison is difficult. For example, which of the four 36 h lead time joint distributions in Figs. 2b–5b is the "best"?

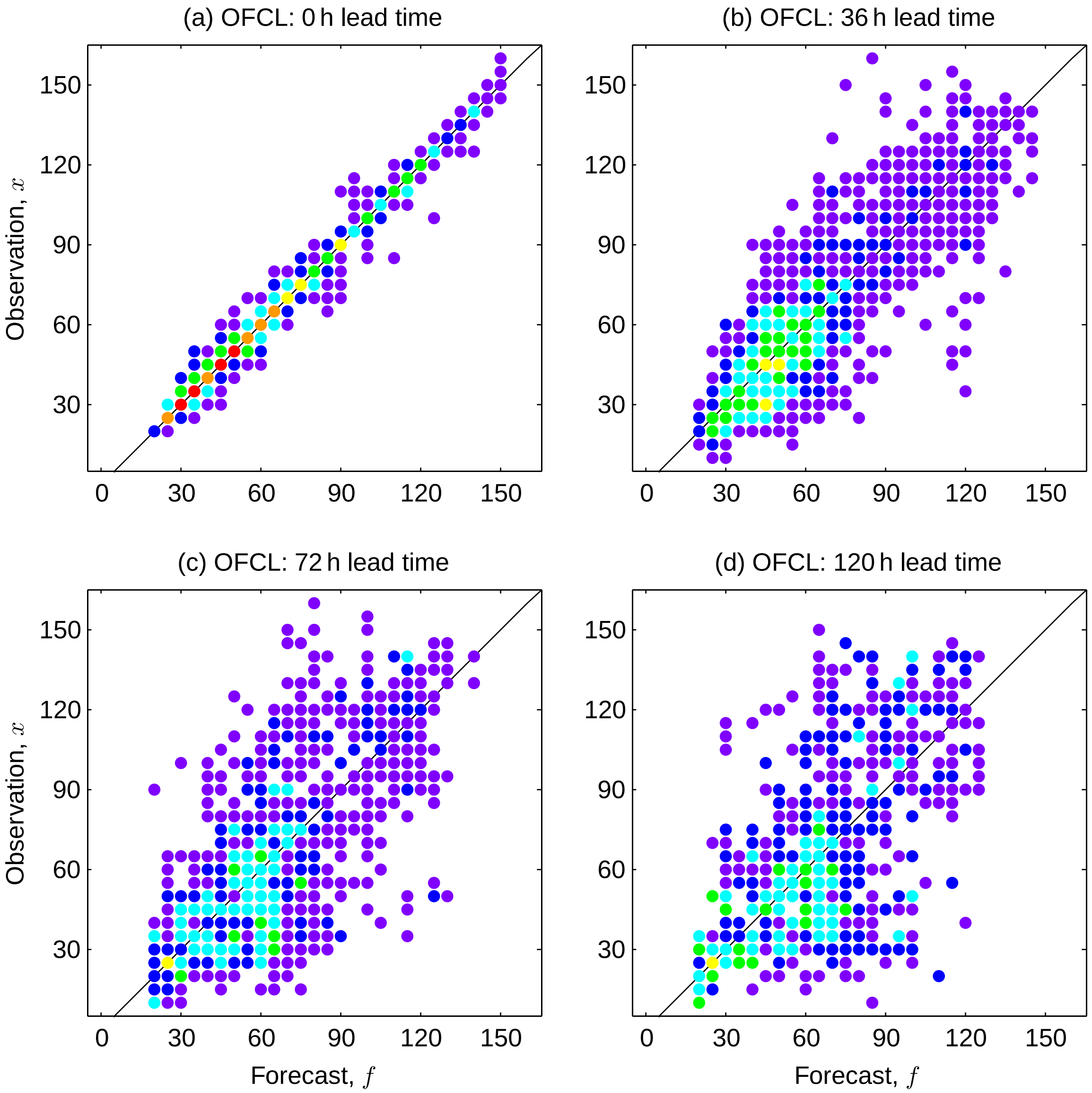

Figure 2: Joint distribution of official NHC intensity forecasts and best track observations at lead times of (a) 0 h, (b) 36 h, (c) 72 h, and (d) 120 h. Dots mark all (f, x) for which there is non-zero relative frequency in the corresponding verification data sample. The colors represent relative frequency magnitude, according to the following scale: $0 < p(f, x) \leq 0.0025$, purple; $0.0025 < p(f, x) \leq 0.005$, dark blue; $0.005 < p(f, x) \leq 0.01$, light blue; $0.01 < p(f, x) \leq 0.015$, green; $0.015 < p(f, x) \leq 0.025$, yellow; $0.025 < p(f, x) \leq 0.05$, orange; $0.05 < p(f, x) \leq 1$, red. The thin black line marks the diagonal, where $f = x$.

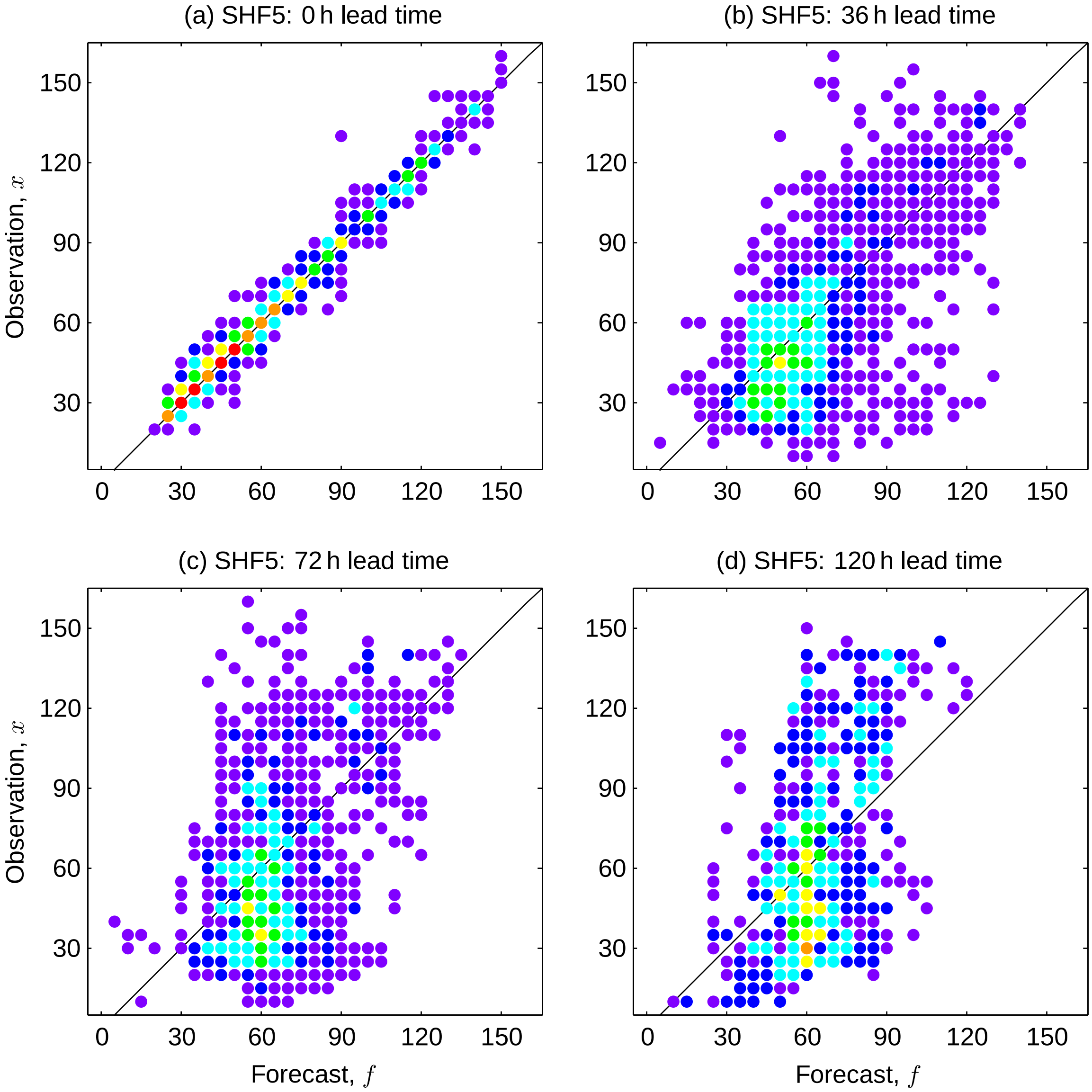

Figure 3: As in Fig. 2, but for SHF5 model intensity forecasts.

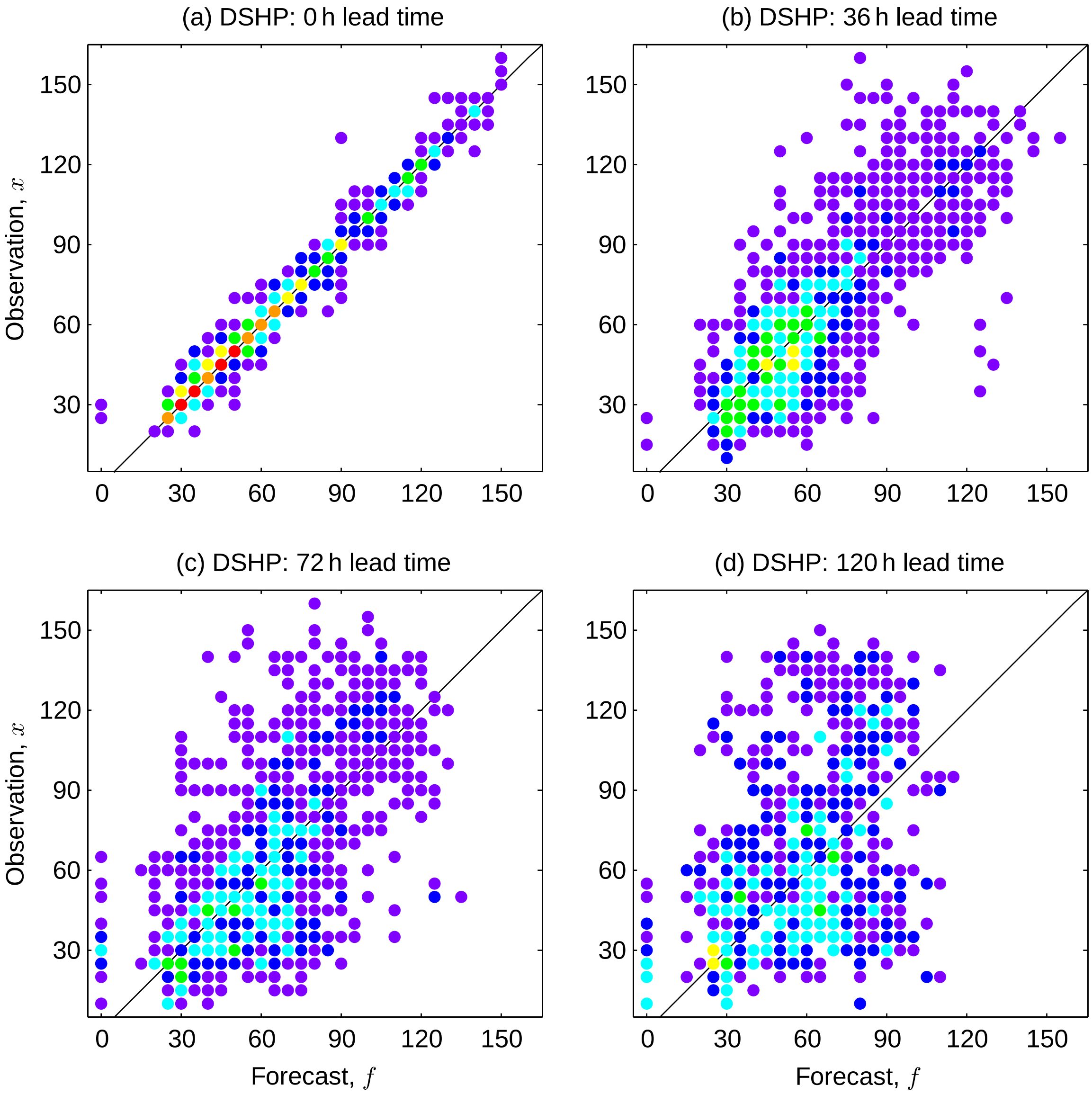

Figure 4: As in Fig. 2, but for the DSHP model intensity forecasts.

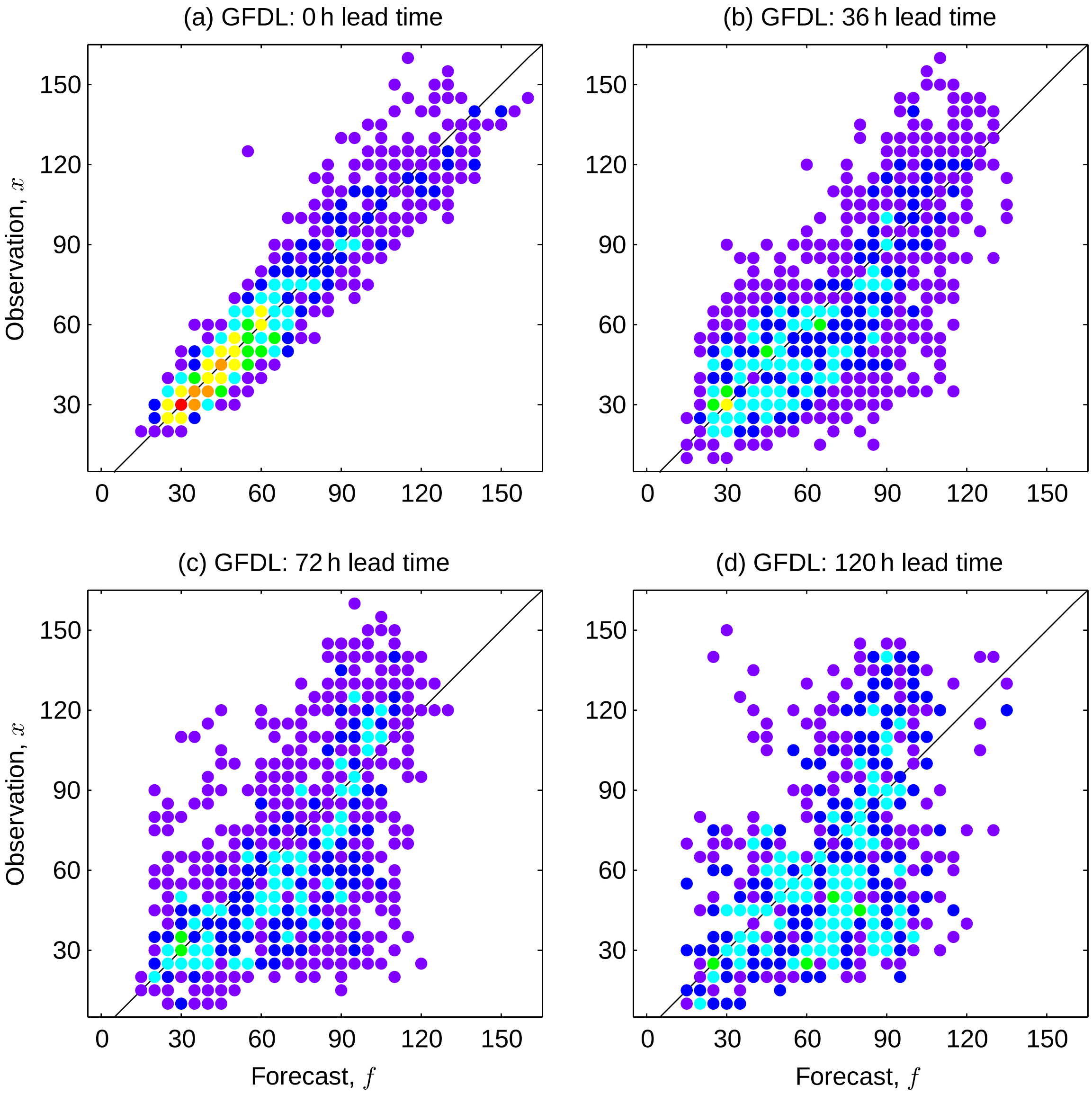

Figure 5: As in Fig. 2, but for the GFDL model intensity forecasts.

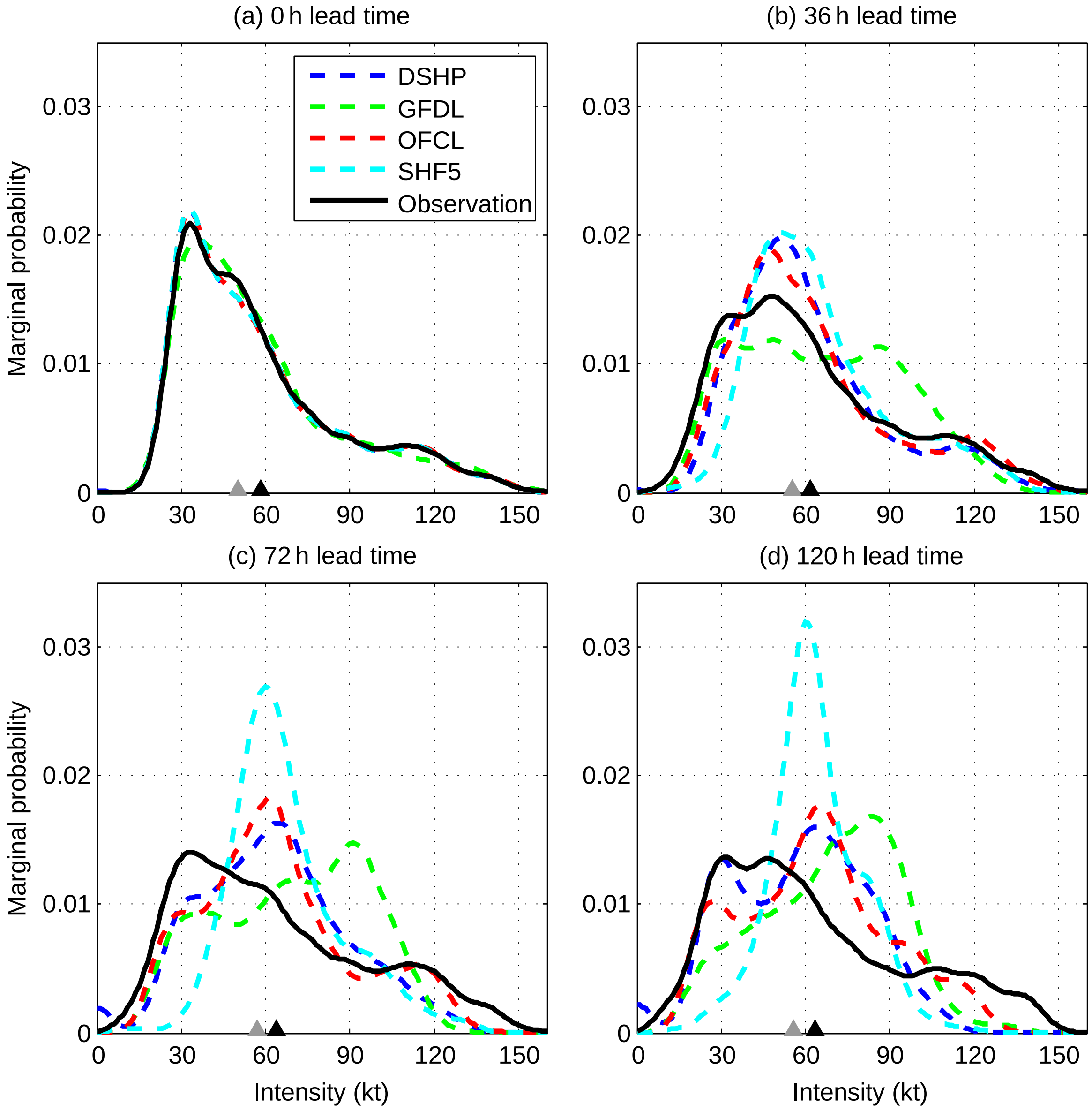

Figure 6: Marginal distributions of OFCL forecasts (dashed red), SHF5 forecasts (dashed light blue), DSHP forecasts (dashed dark blue), GFDL forecasts (dashed green), and observations (solid black) at lead times of (a) 0 h, (b) 36 h, (c) 72 h, and (d) 120 h. The black triangle marks the mean observation and the gray triangle marks the median observation in each panel.