Joint IMD-WMO Group Fellowship Training

on

Numerical Weather Prediction

managed by

India Meteorological Department (IMD), Pune

04 Oct – 10 Nov 2021

Lecture Notes on "*Verification of NWP Forecasts*"

by

Kondapalli Niranjan Kumar

National Centre for Medium Range Weather Forecasting
Ministry of Earth Sciences, Govt. of India
A-50, Sector-62, Noida, UP, India

II. Forecast Verification

Forecast verification involves evaluating the quality of forecasts. Various methods exist to accomplish this. In all cases, the process entails comparing model-predicted variables with observations of those variables. Here, we first focus on how to evaluate the quality of the model forecasts. Most models are under continuous development, and the only way modelers can know if routine system changes, upgrades, or bug fixes improve the forecast quality is to calculate error statistics objectively and quantitatively. It is useful to define some basic terminology such as **accuracy, skill, bias** that we will be employing.

Forecast accuracy (or quality) – a qualitative or quantitative measure of the extent to which a forecast simulates observed conditions. In other words, A measure of the average degree to which pairs of forecast values and observed values correspond. Scalar measures of accuracy summarize the overall quality of the forecasts in the form of a single number.

Forecast skill – a quantification of the improvement in forecast accuracy relative to a simpler baseline method requiring no particular expertise. In other words, The accuracy of a forecast relative to a reference forecast.

Bias – A measure of the correspondence between the average of a forecast variable and the average of the observations.

1. Some standard metrics used for model verification

A. Accuracy measures for continuous variables

These measures apply to variables that are continuous in the sense that they can take on any value within a physically realistic range. In the following definitions F signifies a forecast, O an observation, A an analysis, C climatology and – (an over-bar) the mean

a. Mean Absolute Error (MAE)

$$MAE = \sum_{i=1}^{n} \overline{|F_i - O_i|}$$

Range 0 to∞; Perfect Score = 0

The MAE is the arithmetic average of the absolute difference between pairs of forecast and observed quantities. Cancellation of errors of different sign is avoided by considering the MAE. Errors with different magnitude are given equal weighting.

b. Root Mean Square Error (RMSE)

$$RMSE = \sum_{i=1}^{n} \sqrt{\overline{(F_i - O_i)^2}}$$

Range 0 to ∞; Perfect Score = 0

The RMSE, based as it is on the square of the model error, penalises large errors more heavily and, at a practical level, this may be argued to be a more appropriate way to weight errors compared with the MAE approach. This has the same physical dimensions as the forecast and observations. The above metrics represent both systematic and random components to the error.

c. Anomaly Correlation Coefficient (ACC)

$$ACC = \frac{\overline{(F-C)(A-C)}}{\overline{(F-C)^2(A-C)^2}}$$

Range −100 to 100%; Perfect Score = 100%

An additional, commonly used measure of correspondence between observations and forecasts is the Anomaly Correlation Coefficient (ACC). As the name implies, it is designed to define similarities in the patterns of the departures (i.e., anomalies) of the observed and forecast variables from the climatological means. The ACC can be calculated based on time series or spatial fields and is designed to reward for good forecasts of the pattern (phase and amplitude) of the observed variable. As such the ACC is a positively orientated skill score, meaning that high values are good and that this metric already accounts for climatology and, therefore, quantifies added value. An ACC of 50% is equivalent to a baseline climatology performance, while 60% and 80% represent skilful model performances at the largest weather pattern scales and at the synoptic scales respectively.

The **bias** is the same as the **Mean Error (ME)**, such that

$$ME = Bias = \bar{F} - \bar{O}$$

This is also known as the **systematic error**. Given that $\bar{O}$ is a simple way of defining the climatology of the variable (at least for the limited period of the verification), and $\bar{F}$ is the model climatology for the variable, the bias represents a comparison of the model and actual climatological values.

B. Accuracy measures for discrete variables

These measures apply when the verification question is defined in terms of a yes–no condition. For example, consider a precipitation forecast of whether the accumulated amount is above a specified threshold at a particular location. The observation at that point defines whether precipitation of that amount indeed occurred, a yes or no condition, and the forecast is also in the form of a yes or no. This problem can be illustrated with a 2 × 2 contingency table of the form shown below Fig. 1a.

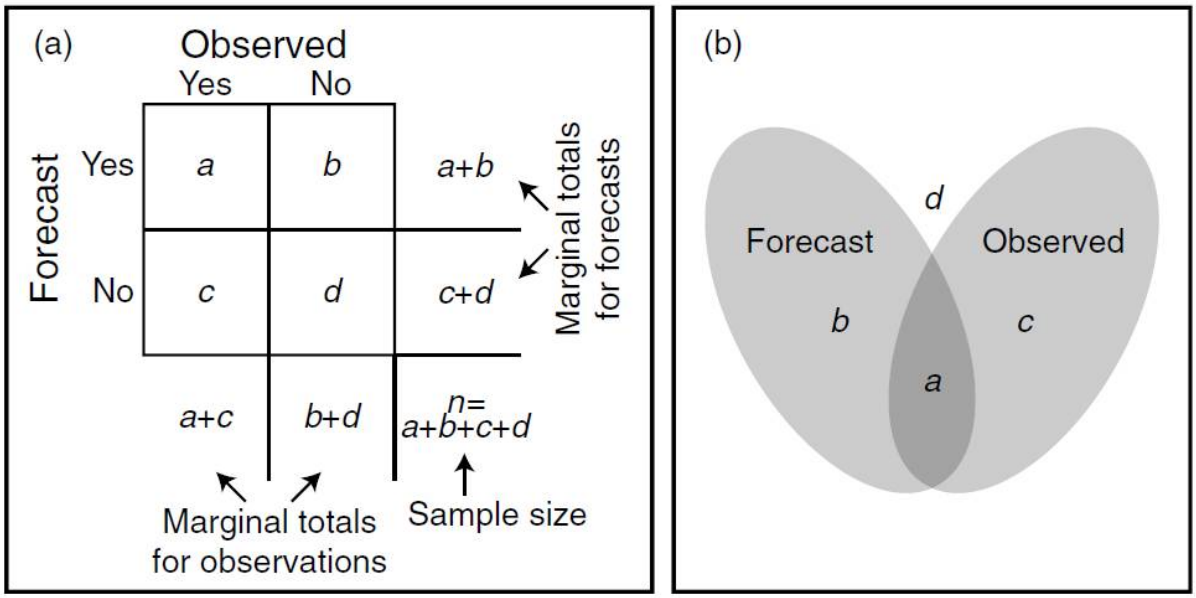

Figure 1: *Contingency table showing the four possible outcomes of a forecast of a discrete variable (a). Also shown is a schematic example of the observed and predicted areas where a variable (e.g., accumulated precipitation) exceeds a specific threshold.*

Of the n forecast–observation pairs, **a** represents the number of times that an observed event was correctly forecast (called hits), **b** is the number of times that no event occurred but the forecast was for an occurrence (called false alarms), **c** is the number of times that an observed event is forecast to not occur (called misses), and d is the number of times that an event was correctly forecast to not occur (called a correct negative). An example is shown in Fig. 1b of areas where a condition (e.g., 24-h accumulated precipitation above a threshold) is observed and forecast. Each area is defined in terms of the elements of the contingency table.

From the contingency table, a number of performance measures can be defined

a. Frequency Bias

$$Frequency\ Bias, FBIAS = \frac{(a + b)}{(a + c)}$$

Frequency Bias quantifies whether the forecast system tends to predict the particular event in question more or less often than it is actually observed, with values less than one indicating that the forecast system tends to underpredict the occurrence and values greater than one indicating a tendency to over-predict occurrence.

b. Hit Rate (Probability of Detection)

$$Hit\ Rate\ (Probability\ of\ Detection), HR(POD)\ =\ \frac{a}{(a+c)}$$

The Hit Rate quantifies the fractional success of the forecast system in predicting the particular event on the occasions when it actually occurs, with a score of one indicating a perfect Hit Rate.

c. False Alarm Rate (FAR)

$$False\ Alarm\ Rate(FAR)\ =\ \frac{b}{(b+d)}$$

The FAR is the ratio of false alarms to the total number of nonoccurrences of the event. Please note that it is different from the False Alarm Ratio, which is the fraction of yes forecasts that are wrong, and is defined as

$$False\ Alarm\ Ratio\ =\ \frac{b}{(a+b)}$$

d. Relative Operating Characteristic (ROC)

The Relative Operating Characteristic (ROC) score is a concept taken from signal processing. It combines Hit Rate (HR) and False Alarm Rate (FAR) by plotting them on an x–y diagram with FAR on the x-axis and HR on the y-axis. A set of perfect forecasts would appear as a point in the top left corner of the diagram – that is FAR = 0 and HR = 1. It means that better forecasts have a low FAR and a high HR, so more-accurate ones have points in the upper-left. The area under the curve has a maximum possible value of unity, corresponding to a perfect forecast. The diagonal corresponds to an unskilled forecast, and the associated area would be 0.5. Forecasts with a degree of skill (i.e. better than random) would appear as points above the diagonal from bottom left to top right, and forecasts with negative skill (i.e. worse than random) would appear as a point below this diagonal. Forecasts with ROC areas of ~0.75 or higher are considered to be good. An example of ROC curve is shown below

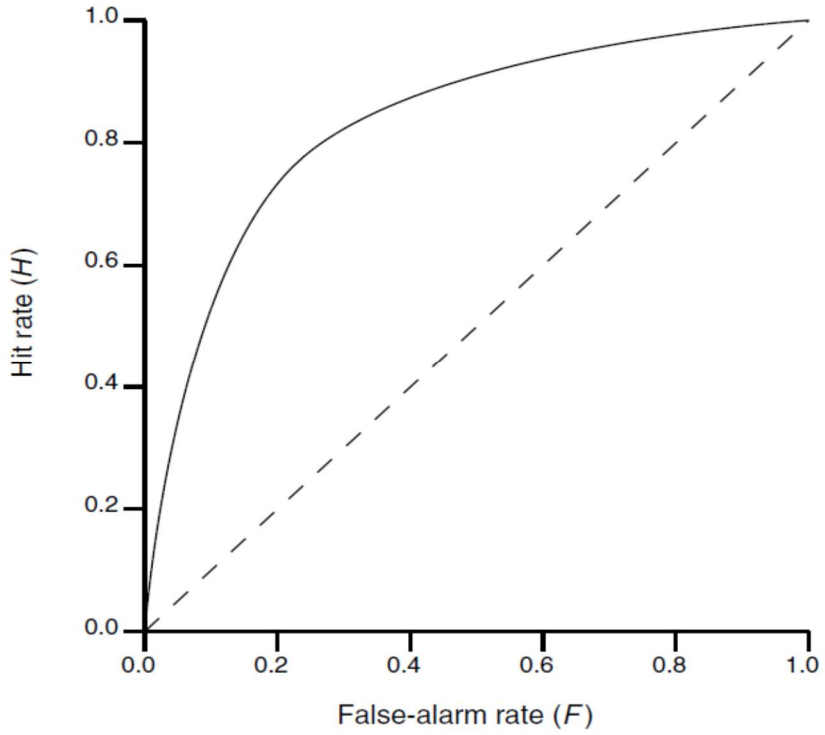

Figure 2: *An example Relative Operating Characteristic (ROC) diagram. The solid line is a ROC curve for a good forecast, the dashed line defines a random forecast, and the area under the ROC curve (fraction of 1.0) represents the overall performance in terms of this metric.*

e. Skill Scores

As noted earlier, skill is defined as the accuracy of one forecast method relative to that of a reference forecast. The skill is usually represented as a Skill Score (SS), which is defined as a percentage improvement over the reference forecast. Mathematically, a SS can be defined as

$$SS_{ref} = \frac{A - A_{ref}}{A_{perf} - A_{ref}} \times 100\,\%,$$

where A is the accuracy of a forecast, A_{ref} is the accuracy of a reference forecast, and A_{perf} is the accuracy of a perfect forecast. If, $A = A_{perf}$ the skill score is 100%. If, $A = A_{ref}$ the skill is zero, indicating no improvement relative to the reference forecast. If the forecast accuracy is less than that of the reference forecast, the skill score is negative. A number of skill scores are based on the previously described 2 × 2 verification contingency table, and have the form of above equation.

f. Heidke Skill Score (HSS)

One of the most-frequently used is called the Heidke Skill Score (HSS) and is based on the proportion correct (PC) as the accuracy measure (A) . The PC here is defined as

$$PC = \frac{(a + d)}{n}$$

(*represents the fraction of the forecasts that correctly anticipated the event or nonevent*)

The reference accuracy measure, A_{ref} , is the proportion-correct value that would be obtained by random forecasts that are statistically independent of the observations. The expression for the HSS is

$$HSS = \frac{2(ad - bc)}{(a + c)(c + d) + (a + b)(b + d)}$$

g. Gilbert Skill Score (GSS) or the Equitable Threat Score (ETS)

Equitable scores are particularly valuable in the verification of deterministic forecasts, since they heavily penalise constant and purely random forecasts. The Threat Score (TS) or Critical Success Index(CSI) is used as the basic accuracy measure(A), and the TS for random forecasts is used as the reference(A_{ref}). The TS can be defined here as

$$TS = CSI = \frac{a}{(a + b + c)}$$

(*which is useful when the yes-event to be forecast occurs much less frequently than the no-event)*

The Gilbert Skill Score (GSS) or the Equitable Threat Score (ETS), and is derived in Wilks (2006) as

$GSS = ETS = \frac{a - a_{ref}}{(a - a_{ref} + b + c)}$, where

$$a_{ref} = \frac{(a + b)(a + c)}{n}$$

The commonly seen **Peirce's Skill Score**, on the other hand, is a truly equitable measure:

$$Peirce's\ Skill\ Score, PSS = \frac{(ad - bc}{(a + c)(b + d)} = POD - FAR$$

Despite the nineteenth century origin of the contingency table, research is still continuing to identify new skill scores for particular applications. For example, although PSS is truly equitable, it is not helpful when evaluating some precipitation forecasts because it is restricted to two categories when, in practice, we might wish to separate dry forecasts from a number of

other amount categories. Many more skill scores, with various strengths and weaknesses, are described in Wilks (2006) and Gilleland et al. (2009).

C. WMO verification metrics

The World Meteorological Organization (WMO) publishes regular verification statistics for all Weather Forecasting Centres which run global NWP model. The categories that the WMO uses for its verification metrics are:

- **Region:** Northern hemisphere extra-tropics (20◦N–90◦N), Southern hemisphere extra-tropics (20◦S–90◦S) and the tropics (20◦N–20◦S).
- **Variable:** pressure, geopotential height, temperature and winds.
- **Level:** mean sea level for pressure; 500 hPa and 250 hPa for geopotential height, winds and temperature in the extra-tropics; and 850 hPa and 250 hPa for geopotential height, winds and temperature in the tropics.
- **Forecast lead time:** T+24 and every 24 hours thereafter.

These metrics give a very broad indication of model performance over wide regions.

D. Verification of probability forecasts

The verification of probability forecasts calls for sophisticated techniques to determine the usefulness or skill of the forecasts.

a. Reliability diagrams

Reliability is an important attribute of ensemble forecasts of dichotomous events – ones that either occur or do not occur at a grid point or over an area – and reliability graphs are a device for easily visualizing the quality of probabilistic forecasts. For each probability bin used in the forecasts, the observed frequency of the particular event being forecast can be calculated from a long record of events. The observed frequency can be plotted against the forecast probability to give a graphical representation of the forecast performance. An example of such a reliability diagram is shown in Figure 3. The best possible set of forecasts would lie perfectly on the bottom-left to top-right diagonal of this diagram (the straight solid line), indicating that the forecasts realistically represented the observed frequency in all probability categories. So, **better forecasts** are closer to the **diagonal line** and worse ones are farther away. The distance of each point from the diagonal gives the **conditional bias**. Points that lie **below the diagonal line indicate over-forecasting**; in other words, the forecast probabilities are too large. The forecast probabilities are too low when the points lie above the line. The reliability diagram is conditioned on the forecasts, so it is often used in combination with the ROC (Figure 2), which

is conditioned on the observations, to provide a “complete” representation of the performance of probabilistic forecasts.

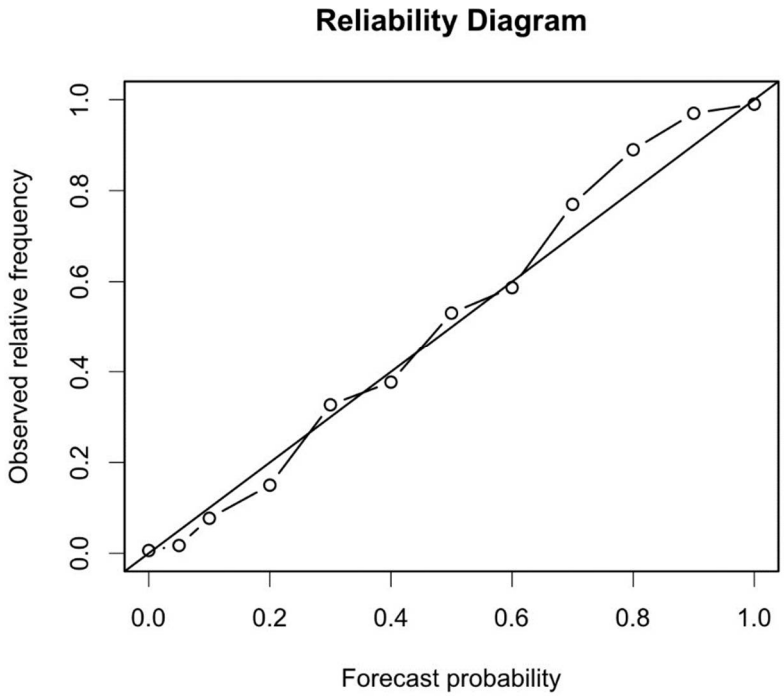

Figure 3: *Example of Reliability Diagram*

The area between the forecast curve (dashed line connected with ‘o’ in Figure 3) and the 1:1 diagonal (the solid line in the Figure 3) can be used as a quantitative measure of the reliability of probability forecasts, with a low value indicating reliable forecasts. The reliability can be written down in the form of an equation as:

$$Reliability = \frac{1}{N}\sum_{i=1}^{N} n_i(f_i - o_i)^2$$

where N is the total number of forecasts, n_i is the number of forecasts in each probability bin, f_i is the forecast probability and o_i is the observed frequency of the event when forecast with probability f_i. A perfectly reliable forecast would have a score of zero.

b. Brier scores and Brier skill scores

The Brier Skill Score (BSS) is based on the Brier Score (BS), which assesses the accuracy of probabilistic predictions. The BS is defined as

$$BS = \frac{1}{n}\sum_{i=1}^{n} (p_i - o_i)^2,$$

And calculates the average squared difference between forecast probabilities (p_i) and observational outcomes (o_i), for n forecast-event pairs, where o is zero if the event does not

occur and unity if it does occur. The BS ranges from zero to one, with lower values indicating better forecasts. The BSS is defined as

$$BSS = \frac{BS - BS_{ref}}{BS_{perf} - BS_{ref}} = 1 - \frac{BS}{BS_{ref}}$$

where $BS_{perf} = 0$. The BSS is unity for a perfect forecast, and zero or negative for unskillful forecasts relative to the reference forecast.

E. Feature-based, event-based, or object-based verification

The most useful information in weather forecasts is often related to changes or events, such as abrupt shifts in temperature or wind speed associated with frontal passages. Thus, model forecast verification can be especially meaningful if it is performed in terms of how well events are forecast. The terms objects, features, or events are used interchangeably in the literature.

For instance, a set of events is defined in a time series of observations, where $o_t - o_{t-2\Delta t}$ is defined as the event in the observations, and $\Delta t = 1hr(could\ be\ day, month, etc.)$. For each observed event, the following quantity is calculated,

$$\frac{\sigma_o}{\sigma_x}\left(\frac{x_t - x_{t-2\Delta t}}{o_t - o_{t-2\Delta t}}\right),$$

Where, $x_t - x_{t-2\Delta t}$ is the change in the model solution for the location and time period of the observed event. The individual observed and forecast event magnitudes are normalized by the respective variances in the two-time series for each location. This ratio is calculated for each station, where the value is +1 for a perfect forecast.

Feature Based Method

The verification of convective precipitation is well known to be especially problematic, and it lends itself to the use of feature-based methods. Other approaches in which analyses and forecasts of precipitation fields are overlaid, and the overlap regions used to compute scores sometimes do not adequately represent the accuracy or value of a forecast. Thus, alternative feature-based approaches have been developed that provide better metrics. An example of this feature-based approach to estimate the statistics is shown in Figure 4.

In summary, the general approach involves

(1) identifying features in the observed (e.g., radar-based analyses) and forecast precipitation fields using thresholds of precipitation amount.

(2) describing the geometric properties of the features (e.g., number, location, shape, orientation, size, average precipitation intensity in the feature).
(3) comparing the relative attributes of the observed and forecast features; and
(4) associating features in the forecast and observed fields, where possible.

Figure 4 illustrates the benefit of this method for verifying precipitation forecasts.

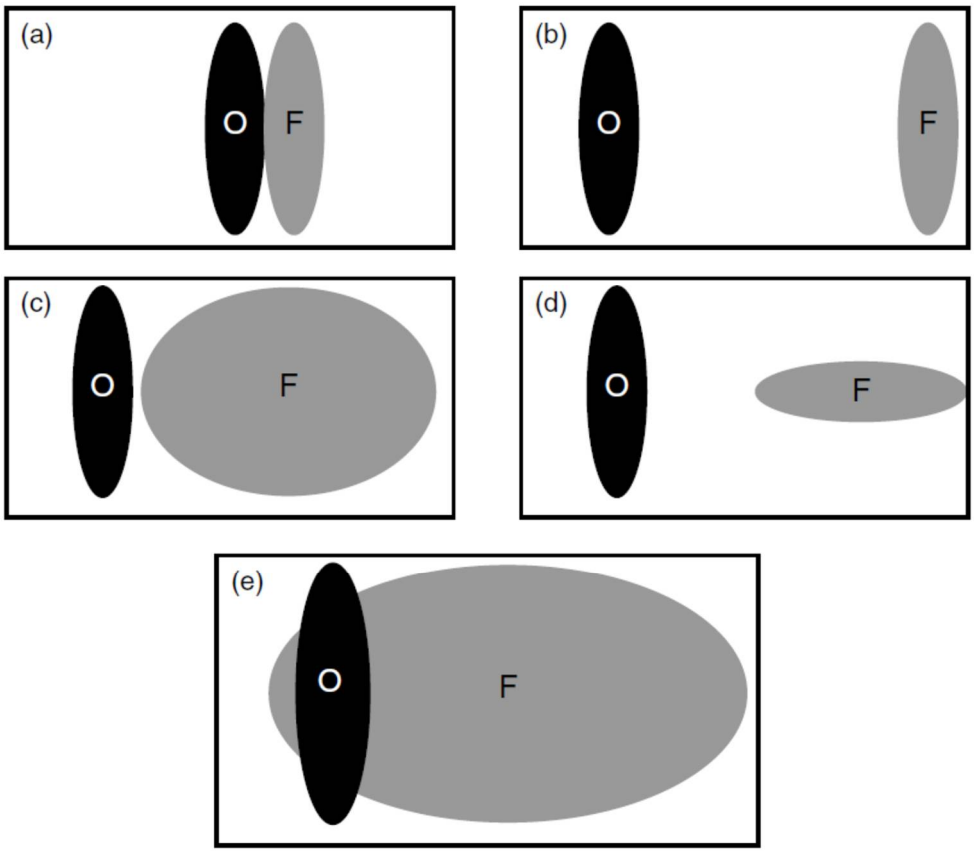

Figure 4: *Schematic example of different combinations of forecasts and observations. From Davis et al. (2006a).*

F. Verification in terms of the scales of atmospheric features

A model solution should approximately preserve the observed spatial and temporal spectra of the dependent variables. Thus, the spectral power for the atmosphere and the model solution have been compared in numerous studies. Of course, unlike other verification criteria, model-simulated features do not need to be in phase with observed features in order for the model to verify well in this context. Rather, the model solution simply needs to contain the features on the correct scales. This type of verification can:

- help the modeler better understand the explicit and implicit spatial and temporal filters in a model

- provide information about whether the lower boundary forcing in the model is imparting the correct scales of motion in the lower troposphere and boundary layer;
- define the model's true resolution.
- illustrate the amount of fine-scale information that is contained in the initial conditions, provided by a data-assimilation system; and
- define the time required for the model to spin up scales of motion that are not in the initial conditions.

An example of spatial verification (Figure 5) related to the typical Kinetic energy spectra also confirm the degree to which the model is faithful to the dynamics of the atmosphere. The sloping straight line is the anticipated spectrum, where the slope depends on the wavenumber range. In general, global-scale models should reproduce the large-scale, k^{-3} slope of the spectrum. In mesoscale and cloud-scale model solutions, the slope should be $k^{-5/3}$. When model resolutions span the global scale and the mesoscale, as is the case with high-resolution global models, the verification of the existence of the slope transition in the kinetic-energy spectrum is a test that the model is faithful to the atmospheric dynamics. And analyses of kinetic-energy spectra have been used to verify the ability of a model to represent scales near the $2\Delta x$ limit of resolution. This latter type of verification defines the effective resolution of the model.

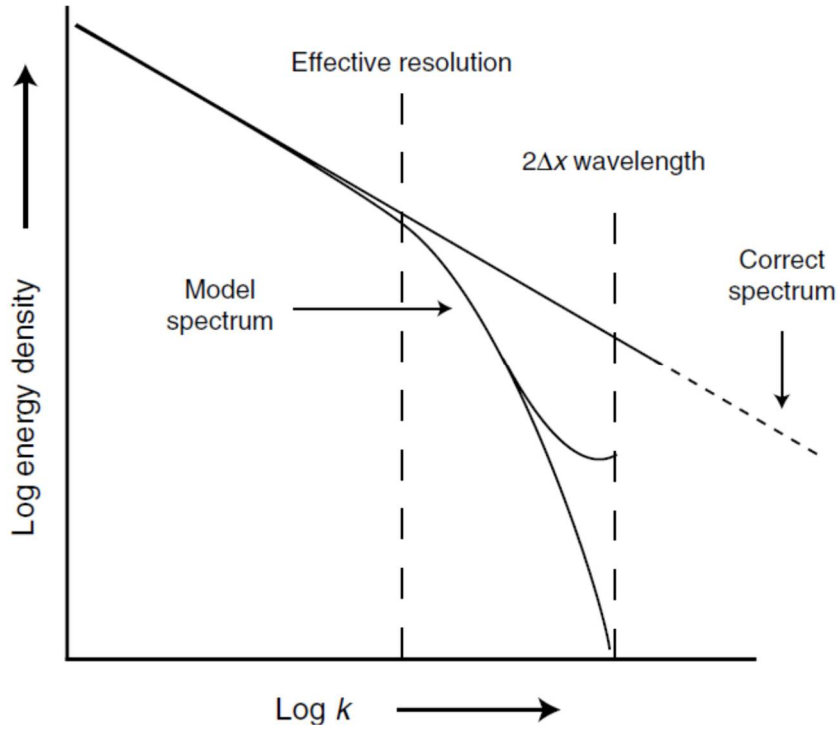

Figure 5: *Schematic of the theoretical and model kinetic-energy spectrum, showing the effective resolution. Two examples are shown of the decay in the kinetic energy at the high-wavenumber end of the model spectrum. Adapted from Skamarock (2004).*

Summary

- The verification of forecasts is an important aspect of the work of a forecasting centre, feeding into the development and improvement of forecast models.
- Verification is also important to users who can make informed choices based on the skill of different forecast providers and can also modify their weather-dependent behaviour based on the known characteristics of the forecasts they are using.
- The World Meteorological Organization (WMO) oversees a verification of the global NWP models from all the major forecasting centres in the world, using a range of different domains, forecast variables and lead times.
- Forecast skill measures the ability of a particular forecasting method to improve upon forecasts made using some zero-skill technique such as persistence or climatological forecasts.
- For simple deterministic forecasts predicting the occurrence or nonoccurrence of a particular meteorological condition, contingency tables can generate a wide range of verification scores which can be of particular relevance to forecast customers.
- Verification of probability forecasts is complex, but the use of reliability diagrams and the Brier Score can provide useful quantification of forecast performance.
- Probability forecasts can help customers to make economic decisions on the basis of the weather forecast.

SUGGESTED GENERAL REFERENCES FOR FURTHER READING

1. Wilks, D. S. (2006). Statistical Methods in the Atmospheric Sciences. San Diego, USA: Academic Press.
2. T. Warner, Numerical Weather and Climate Prediction (Cambridge Press, Cambridge, UK, 2011)
3. P. Inness, S. Dorling, Operational Weather Forecasting, DOI:10.1002/9781118447659 (John Wiley & Sons, Ltd)
4. Haraldur Ólafsson, Jian-Wen Bao (Eds), Uncertainties in Numerical Weather Prediction, Elsevier, 2021, https://doi.org/10.1016/B978-0-12-815491-5.09991-2.
5. Davis, C., B. Brown, and R. Bullock (2006a). Object-based verification of precipitation
6. forecasts, Part I: Methodology and application to mesoscale rain areas. Mon. Wea. Rev.,134, 1772–1784.
7. Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert (2009). Intercomparison of spatial forecast verification metrics. Wea. Forecasting, 24, 1416–1430.
8. Skamarock, W. C. (2004). Evaluating mesoscale NWP models using kinetic energy spectra. Mon. Wea. Rev., 132, 3019–3032.
9. Website: https://dtcenter.org/community-code/model-evaluation-tools-met