Chapter 7
Verification and Evaluation of Load Forecast Models

What makes a good forecast? This section aims to introduce some of the main tools for evaluating the quality of time series forecasts. It is worth noting that this is still a very active research area, especially in the developing area of probabilistic load forecasts. Obviously error measures can only be calculated after the actual observations have become available, although in practice forecasts are evaluated on the historical data by splitting it into training and testing periods (see Sect. 8.1.3).

Of course, when a forecast is required for a particular application why is it not more appropriate to simply evaluate the forecast based on its performance for that application? One of the reason's is that the performance of an application (See examples of applications in Chap. 15) is not usually defined in a simple way and may be computationally infeasible, especially if multiple evaluations are required. Instead simpler, easier to calculate measures, such as those introduced in this chapter are used. However, it still does not mean that any measure can be used and it is always preferable that one is chosen which aligns to the application as closely as possible.

This section will begin by introducing error measures which are used for both evaluating the accuracy of the forecasts but are also to compare and select between various models (Sect. 8.2). Before looking at specific error metrics and measures it is worth noting that the measures have to be different depending on whether we are considering point, or probabilistic forecasts (Sect. 5.2) with the latter having several different forms which may require different measures. The next section considers point forecast measures, and then probabilistic error measures are discussed in Sect. 7.2. These measures can be used to define skill scores, an important evaluation method for forecast skill, and are considered in Sect. 7.4. The chapter then finishes by illustrating ways to improve a forecast based on residual checks and other forecast correction methods.

S. Haben et al., *Core Concepts and Methods in Load Forecasting*,
https://doi.org/10.1007/978-3-031-27852-5_7

7.1 Point Forecast Error Measures

To define the error measures, consider two h-step-ahead point forecasts

$$\hat{\mathbf{L}}^{(1)} = (\hat{L}^{(1)}_{n+1}, \hat{L}^{(1)}_{n+2}, \ldots, \hat{L}^{(1)}_{n+h})$$

and

$$\hat{\mathbf{L}}^{(2)} = (\hat{L}^{(2)}_{n+1}, \hat{L}^{(2)}_{n+2}, \ldots, \hat{L}^{(2)}_{n+h})$$

for a time series with actual values given by $\mathbf{L} = (L_{n+1}, L_{n+2}, \ldots, L_{n+h})$. The errors between the forecasts are defined by

$$\mathbf{e}^{(k)} = \mathbf{L} - \hat{\mathbf{L}}^{(k)} = (L_{n+1} - \hat{L}^{(k)}_{n+1}, L_{n+2} - \hat{L}^{(k)}_{n+2}, \ldots, L_{n+h} - \hat{L}^{(k)}_{n+h}) = (e^{(k)}_1, e^{(k)}_2, \ldots, e^{(k)}_h), \tag{7.42}$$

where k is 1 or 2. How can these forecasts be scored and these errors summarised in order to compare which one is 'closer' to the actual values and hence which is more accurate? The answer is not obvious as there are several ways to choose how to measure this (unless $\mathbf{e}^{(k)} = \mathbf{0}$ of course, in which case you've achieved a perfect forecast!).

As an initial choice, consider norm functions, a common way of measuring the distance between vectors. Given any real-valued vector $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, the p-norm of $\mathbf{x}$ is defined to be

$$||\mathbf{x}||_p = \left(\sum_{k=1}^{N} x_k^p \right)^{1/p} = (x_1^p + x_2^p + \cdots + x_N^p)^{1/p}, \tag{7.43}$$

where $p \geq 1$. The most common norms are the 1-norm (i.e. the absolute sum),

$$||\mathbf{x}||_1 = |x|_1 + |x|_2 + \cdots + |x|_N, \tag{7.44}$$

the 2-norm (known as the standard Euclidean norm),

$$||\mathbf{x}||_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_N^2}, \tag{7.45}$$

and also the ∞-norm which is defined as $||\mathbf{x}||_\infty = \max_{k \in \{1,\ldots,N\}} |x_k|$. The p-norms are **metric functions** which have the following useful properties which make them well-defined and intuitive for measuring the difference between two vectors:

1. **Positive Definite**: $||\mathbf{x}||_p \geq 0$ and $||\mathbf{x}||_p = 0$ if and only if $\mathbf{x} = \mathbf{0} = (0, 0, \ldots, 0)$. In other words the sizes are always positive and only zero if all the elements of the vector have no size.
2. **Triangle Inequality**: For two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$ then $||\mathbf{x} + \mathbf{y}||_p \leq ||\mathbf{x}||_p + ||\mathbf{y}||_p$. This has the intuitive interpretation that the distance from A to B and then B to C will always be longer than the distance directly from A to C.

Table 7.1 A comparison of different p-norm values applied to two vectors as described in the main text (values to the nearest 2 decimal places)

p-norm	$\mathbf{e}^{(1)}$	$\mathbf{e}^{(2)}$
1	1.2	1.2
2	1.01	0.69
∞	1	0.4

The choice of p determines the emphasis of the p-norm on different components of the error, with larger p values meaning that the norm more strongly represents the larger error components. To illustrate this consider two error vectors $\mathbf{e}^{(1)} = (1, 0.1, 0.1)$ and $\mathbf{e}^{(2)} = (0.4, 0.4, 0.4)$, produced by two different 3-step ahead forecast models. The first model has a relatively large peak error whilst the second has no such large errors and has constant errors at each time step. The errors scores for each forecast for three p-norms with different values of p are shown in Table 7.1. First, notice that the sum of the absolute errors are equal for both forecasts and hence $||\mathbf{e}^{(k)}||_1 = 1.2$ for $k = 1, 2$. Thus the 1-norm evaluates both forecasts as having the same errors. In contrast the ∞-norm only focuses on the largest value and hence gives values of 1 and 0.4 for forecast model 1 and 2 respectively, but doesn't take into account any information about the other errors. Choosing a value of p between these extremes will produce an error value which includes contributions from all error values but with stronger influences from the larger values the larger the p value. In this example it can be seen that the 2-norm produces a similar value for $\mathbf{e}^{(1)}$ as the ∞-norm but has a value for $\mathbf{e}^{(2)}$ which is between both the 1-norm and ∞-norm. Hence the 2-norm includes a contribution from all components of the error but the larger errors contribute slightly more than the 1-norm. The point of this example is that the choice of error measure is an important aspect of the application being considered.

Despite the potential subjectivity in the choice of error measure there are some common methods which are applied in time series, and in particular load forecasting. On their own norms are not usually appropriate as error measures as they don't scale with the problem. The size of the errors will grow with the length of the series which inhibits the comparison of different forecast horizons. For this reason they are often combined with normalisations. One of the most common error measures is the mean absolute error (MAE), defined using the 1-norm as

$$\text{MAE}(\mathbf{L}, \hat{\mathbf{L}}) = \frac{1}{h}||\mathbf{L} - \hat{\mathbf{L}}||_1 = \frac{1}{h}\sum_{k=1}^{h}|L_{n+k} - \hat{L}_{n+k}|. \tag{7.46}$$

The MAE gives an average absolute error for all time points in the forecast horizon, $n + 1, \ldots, n + h$. A useful property of absolute error measures is that the units of the error are often the same as the data being considered which simplifies interpretations.

Since in much of the examples considered in this section the data will be in energy units (e.g. kWh), the errors will be in the same units as well.

Another common error measure, which also preserves the units, is the root-mean-square error (RMSE) which is defined in terms of the 2-norm as

$$\text{RMSE}(\mathbf{L}, \hat{\mathbf{L}}) = \frac{1}{\sqrt{h}}||\mathbf{L} - \hat{\mathbf{L}}||_2 = \sqrt{\frac{\sum_{k=1}^{h}(L_{n+k} - \hat{L}_{n+k})^2}{h}}. \tag{7.47}$$

As seen earlier in this section, the power on the error measure can play an important role in what type of errors the measure focuses on. The higher the power the more focus the measure has on larger errors. Hence for RMSE, the larger errors will contribute *relatively* more to the overall score than with the MAE. This can be important if you are interested in assessing which forecast may be more suitable in accurately estimating extreme values, such as peaks in demand.

A drawback of absolute-type error measures such as MAE and RMSE is the difficulty in making comparisons of accuracy when time series have different magnitudes. For example, an error of 1 kWh in a day ahead forecast is quite significant when the daily demand is only 2 kWh but negligible for substation feeders which regularly have daily demands of 100 kWh or more. A more accurate comparison of these errors may be to present the percentage errors *relative* to the size of values in the time series. In this case the 1 kWh error is 50% of the overall demand for the substation with 2 kWh daily demand but only 1% for the substation feeder with 100 kWh daily demand.

One of the most commonly used scores for evaluating the relative accuracy of a point forecasts is the mean absolute percentage error (MAPE) defined by

$$\text{MAPE}(\mathbf{L}, \hat{\mathbf{L}}) = \frac{100}{h}\sum_{k=1}^{h}\frac{|L_{n+k} - \hat{L}_{n+k}|}{|L_{n+k}|}. \tag{7.48}$$

The individual error at each time step, $|L_{n+k} - \hat{L}_{n+k}|$, is divided by the absolute demand $|L_{n+k}|$ and averaged to give a relative score. The score is often multipled by 100 in order to provide a percentage score. The MAPE is not appropriate for series which have zero or very small values, for example, household level electricity demand, or on a feeder with a lot of localised generation. Small L_k values will inflate the size of the errors at time t_k, masking the true accuracy of the forecast. The MAPE is not even defined when the true value is zero, $L_k = 0$. One alternative employed in this book is to instead replace the denominator with the average value $\frac{1}{N}\sum L_k$. This is known as the weighted absolute percentage error (WAPE). The same scaling can also be applied to the RMSE and MAE to create relative error measures. Above are some of the most common error measures used in load forecasting but of course there are several other measures which could be used, including those which avoid the issue of dividing by zero.

Directly comparing error measures between forecasts can help compare forecast accuracy but they can be complicated if the underlying time series has varying levels of predictability. In Sect. 7.4 **skill scores** are discussed which are very useful for comparing forecast models by utilising a common benchmark to help with interpretation.

It is important to carefully select the error measure that suits the application or purposes of the forecast. A special case will demonstrate this in Sect. 13.3, which presents household level load forecasts. Many standard error measures (including the ones presented here) may be inappropriate for providing an objective score for evaluating the accuracy of a forecast. Instead a novel approach is considered showing that there is no need to restrict your evaluation methods to the most popular or common ones such as RMSE or MAPE.

7.2 Probabilistic Forecast Error Measures

The above scores are only applicable to point forecasts and are not appropriate for assessing probabilistic forecasts. These forecasts are less straightforward to evaluate due to the increased complexity and the various forms that probabilistic forecasts can take (quantile, density, ensembles, etc.) as described in Sect. 5.2. In this section, the focus will be on scores for univariate[1] probabilistic forecasts. Multivariate scoring functions are only discussed in passing but suggestions for further reading can be found in Appendix D.2.

The aim with a probabilistic forecast is to accurately represent the *distribution* of the variable. Since there is only usually one observation per time step, to assess a probabilistic forecast usually means comparing the single observation against the estimate of the distribution. This makes the situation much more complicated compared to point forecasts, which can compare the single observation to the single point estimate value. Ideally the aim is to use a scoring function for which the *minimum* score is only achieved by the true distribution, these are called **proper scoring functions**.

One of the more popular proper scoring functions is the **pinball loss score** or **quantile score** which measures the accuracy of a quantile forecast (Sect. 5.2). Recall that the quantile $\tau \in [0, 1]$ of a CDF, F, is simply the value z_τ such that $F(z_\tau) = \tau$, or in other words, for a univariate distribution the probability of a random variable being less than z_τ is τ (see also Sect. 3.2 for more details on quantiles). Given an estimate of the τ quantile, z_τ, and an actual observation L from the distribution being estimated, the pinball loss function is given by

$$L_\tau(L, z_\tau) = \begin{cases} \tau(L - z_\tau) & L \geq z_\tau \\ (1 - \tau)(z_\tau - L) & L < z_\tau \end{cases}$$

[1] Recall univariate means a single variable whereas multivariate is more than one.

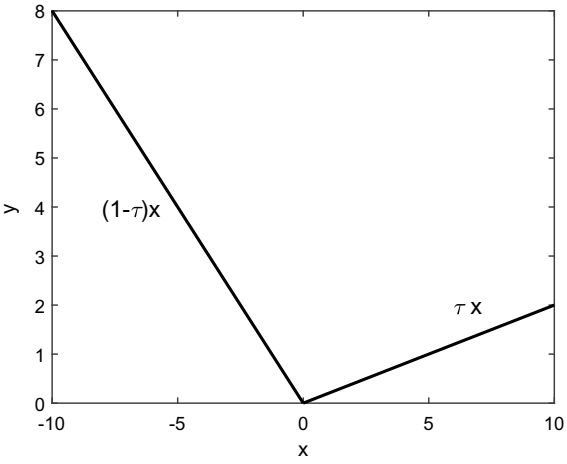

Fig. 7.1 Example of the weighting given by the pinball loss function for $\tau = 0.2$. If the input is positive then the weighting is τ, if negative, then the weighting is $1 - \tau$

The pinball function is an asymmetric function which takes the difference between the observation and the quantile and then weights the difference differently depending on whether the value is positive or negative. This asymmetry is important since an accurately estimated quantile will have, on average, a proportion, τ, of the observations below z_τ. The pinball function and its weighting is illustrate in Fig. 7.1. Typically quantile forecasts are estimated for a series of quantiles z_{τ_k}, for each time step $k = 1, \ldots, N$ in the forecast horizon and these quantiles split the range of the distribution into evenly spaced points $\tau_1, \ldots, \tau_N$ (for example popular values are the deciles $0.1, 0.2, \ldots, 0.9$, or ventiles, $0.05, 0.1, \ldots, 0.9, 0.95$). The **pinball loss score** (PLS) is simply the average over each individual loss over each quantile

$$\text{PLS} = \frac{1}{N} \sum_{k=1}^{N} L_{\tau_k}(L, z_{\tau_k}). \tag{7.49}$$

Another common proper scoring function is the continuous ranked probability score (CRPS). Consider a cumulative distribution $\hat{F}(z)$, which is an estimate of the distribution at some time for which there is an observation, defined as L. The CRPS is defined as

$$\text{CRPS}(L, \hat{F}) = \int_{-\infty}^{\infty} (\hat{F}(z) - \mathbf{1}(z - L))^2 dz = \mathbb{E}(|Z - L|) - \frac{1}{2}\mathbb{E}(|Z - \tilde{Z}|), \tag{7.50}$$

where $\mathbf{1}$ is the Heaviside step function

$$\mathbf{1}(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}. \tag{7.51}$$

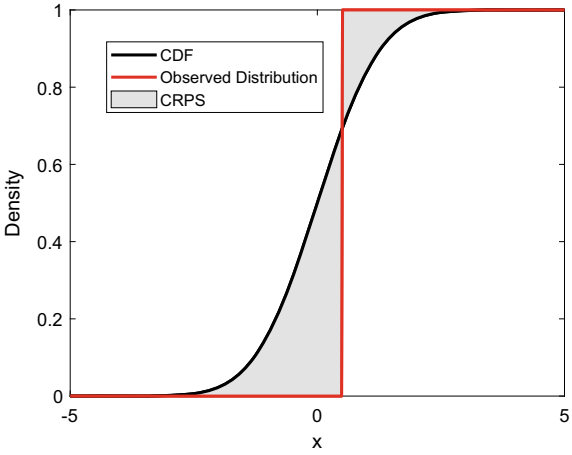

Fig. 7.2 Example of the CRPS which is the area between the CDF and the empirical distribution formed from a single observation

The first, integral form, measures the difference between the estimated distribution $\hat{F}(z)$ and the empirical cumulative distribution function for a single observation, given by $\mathbf{1}(z - L)$. The CRPS for a single observation and the estimated CDF is illustrated in Fig. 7.2, and is equal to the shaded area between the CDF and the empirical distribution (see Sect. 3.4) for the observation. The aim is to minimise this shaded area, and this is achieved by accurately estimating the true distribution.

Notice that the two terms in the second form of the CRPS describe two components of the error. The first term, $\mathbb{E}(|Z - L|)$ is the (expected) absolute difference between the observations and the forecasts. The second term, $\mathbb{E}(|Z - \tilde{Z}|)$, is a measure of the spread, i.e. the *sharpness*, of the probabilistic forecast. For a point forecast the CRPS reduces down to the first term only, i.e. the Mean absolute error. This second equivalent form of the CRPS in Eq. (7.50), $\mathbb{E}(|Z - L|) - \frac{1}{2}\mathbb{E}(|Z - \tilde{Z}|)$ suggests another way of estimating the CRPS using sample means calculated by generating random draws, $\tilde{Z}$ and Z, from the estimated distribution $\hat{F}$. For multiple observations the final CRPS is simply the average of the individual CRPS values.

The CRPS and pinball score are only suitable for univariate densities. For ensemble/scenario forecasts the second form of the CRPS given in Eq. (7.50) can be adapted to cope with ensemble forecasts which estimate a multivariate distribution. Consider a multivariate probability distribution $\mathbf{F_Z}$ which is defined for a N-dimensional random variable $\mathbf{X} = (X_1, X_2, \ldots, X_N)^T$. Given a single observation vector $\mathbf{L} = (L_1, L_2, \ldots, L_N)^T$ then the **energy score** is defined as

$$\text{ES}(\mathbf{L}, \mathbf{F}) = \mathbb{E}(||\mathbf{Z} - \mathbf{L}||_2) - \frac{1}{2}\mathbb{E}(||\mathbf{Z} - \tilde{\mathbf{Z}}||_2), \tag{7.52}$$

where $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ are independent copies of random draws/samples from the multivariate distribution. To calculate this in practice the samples, $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$, are taken from the generated forecast ensembles and the sample means are used to estimate the expected

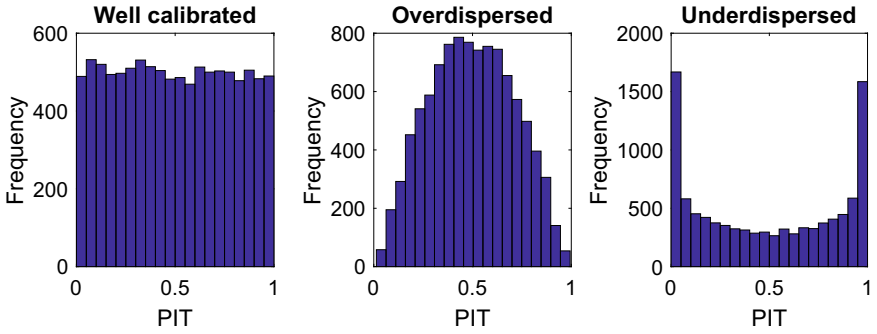

Fig. 7.3 Example of the histograms for the PIT from applying (left) the true Gaussian CDF of mean 2 and standard deviation 0.25, (middle) from applying a Gaussian CDF of mean 2 and standard deviation of 0.4 and (right) from applying a Gaussian CDF of mean 2 and standard deviation equal to 0.15

values. For the pinball score, CRPS and energy score, a smaller value implies a more accurate probabilistic forecast.

Probabilistic forecasts can also be assessed visually. Consider a CDF, F, for a continuous random variable X, then the probability integral transform (PIT) of the data, is defined by the application of the CDF to the observations $p_t = F(X_t)$. The histogram of the PIT should be uniformly distributed if the correct CDF has been chosen. To understand this consider a quantile forecast which estimates the demi-deciles, i.e. the q quantiles where $q = 0.05, 0.1, 0.15, \ldots, 0.9, 0.95$ of a continuous cumulative density function $F(x)$. If the forecast was correctly calibrated 5% of the observations should fall between any consecutive quantiles, $F^{-1}(q)$, and $F^{-1}(q + 0.05)$. In other words, the histogram of the PIT defined by this quantile estimate should be uniform with 5% of observations within each bin.

An example of the PIT histogram is shown in Fig. 7.3 for three different Gaussian CDFs (with different standard deviations) applied to random samples from one of the distributions. When the true CDF is applied (left in the Figure) then the histogram is uniform as expected. When a Gaussian CDF is applied which has a larger standard deviation than the true data then the PIT has too many observations in the centre of the histogram, and the distribution is called overdispersed (middle plot). Alternatively, if a PIT is applied using a Gaussian CDF with a smaller spread (smaller standard deviation) then there is too many observations at the edges of the histogram and the distribution is called underdispersed. Other shapes of the PIT can suggest other biases or inaccuracies in the probabilistic estimate.

An equivalent method for visualising the quality of a probabilistic forecast is a **reliability plot** (or reliability diagram). For this the quantiles of the probabilistic forecast are plotted against the observed relative frequency. In other words, take the τth quantile $F^{-1}(\tau)$, for a probabilistic forecast with CDF, F, and suppose there are observations $y_1, y_2, \ldots, y_N$. These points can be used to define an empirical distribution function $F_E(y)$ which is a step function defined by

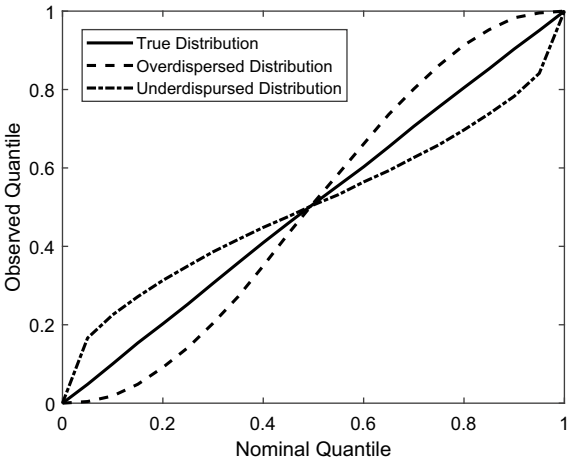

Fig. 7.4 Reliability diagram for the three different estimates for the distribution for the same data as in Fig. 7.3

$$\hat{F}_E(y) = \frac{\text{number observations less than } X}{N} = \frac{1}{N}\sum_{k=1}^{N} \mathbf{1}_{y_k < y}, \tag{7.53}$$

where $\mathbf{1}_S$ is the indicator function which takes the value 1 if the statement S is true and 0 otherwise (also see Sect. 3.4). A reliability diagram is simply a comparison of the quantiles of the estimated distribution, F, with the empirical distribution, F_E. The quantiles should be similar (for the same probability value τ) if the estimate F is an accurate representation of the distribution of the observations (as estimated by the empirical CDF, F_E). The reliability diagram for the same distributions as in Fig. 7.3 are shown in Fig. 7.4. This is for 1000 observations from the true normal distribution with mean 2 and standard deviation 0.25. Notice that in the reliability diagram if the observations are from the true distribution they should be close to the diagonal $y = x$. In contrast the line for the overdispersed distribution has a small gradient for the under-represented tail quantiles but high gradient in the middle for the over-represented quantiles. The opposite is true for the underdispersed distribution which has high gradients at the tail quantiles and a lower gradient in the central quantiles.

It is worth noting, that a uniform PIT (or equivalently a reliability plot lying on the $y = x$ line) is only a necessary condition and not a sufficient condition for the distribution to be the true underlying distribution for the data. In other words, uniform PITs can still occur even if the estimated distribution is not a true representation of the underlying distribution.

7.3 Causes of Forecast Error

There will always be some error in the forecasts no matter the sophistication of the model. However, there are some common causes of forecast error it is worth briefly mentioning to prevent jumping to conclusions and assist in interpreting the models and their errors.

Even when an accurate model has been generated with both low-bias and low-variance the forecasts errors are likely to increase with the forecast horizon, this is illustrated in Fig. 7.5. This is because there is usually a interdependency between values which are close in time. This is particularly true in energy demand behaviour where appliances are used over several hours (heaters), or similar actions are performed together (a morning shower followed by boiling a kettle for a cup of tea). Hence if comparing the errors for forecasting tomorrow, versus forecasts for the following day and so on, there would be an expected upward trajectory in the forecast errors/scores.

However, things may not be as simple as this. In the Case study in Chap. 14, the forecast errors vary within a day (see Fig. 14.7 in particular). This is because there is more variation in demand (and hence larger errors on average) during certain periods of the day compared to others. However, even in this case the average daily errors do seem to be increasing. This highlights another source of forecast error which is the volatility of particular periods.

Another source of forecast error may be due to their dependence on the input variables. Many load forecasts are strongly related to weather (e.g. see Sect. 6.7) and therefore weather forecasts are utilised within the load forecast models. However, if these inputs are inaccurate (for example through measurement, forecasts or even calibration errors) then the load forecast will also be inaccurate. In other words, for surprising or unusual errors in the data, the input variables should also be considered as possible causes.

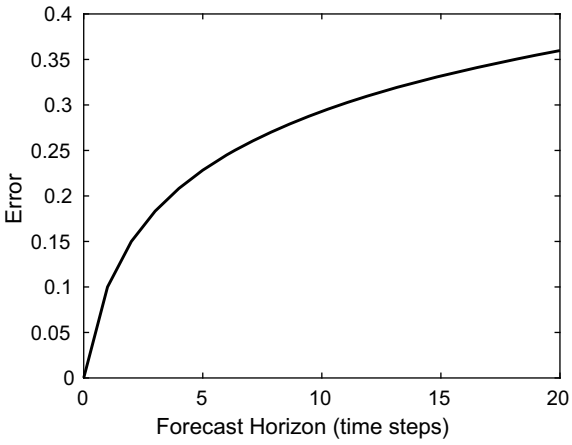

Fig. 7.5 Expected forecast error as a function of forecast horizon

Benchmarks also serve a useful function of determining the causes of forecast errors. Since they may include different inputs than the main models they can confirm which variables may be sources of large errors. Benchmarks are also useful for comparing models and understanding improvements over time, even when the underlying data changes. This is explored in the next section.

7.4 Skill Scores

Even if an error measure is appropriate to the application, it may not be easy to compare or evaluate forecasts, especially if comparing on multiple datasets. For example, consider two forecast models where one model produces an estimate for one dataset and the other model produces an estimate for another, dataset. If these datasets have different volatilities (for example it could be that the data is for different seasons, where say heating appliances may make Winter demand behaviour more volatile) then it will not be clear how to compare the accuracy of these forecasts. Similarly, how do you keep track of the improvement (or degradation) in the same forecast over time, which will be using more and/or newer data?

One way to help discriminate between forecasts in cases like the above and others, is to use a **skill score**. A skill score measures the accuracy of a forecast *relative* to some benchmark score. They are very common in numerical weather prediction applications, where they are used to show the improvement of forecast models over time.

Skill scores can take many forms but a common format is

$$SS(\hat{\mathbf{L}}, \mathbf{L}_b) = \frac{E_f - E_b}{E_p - E_b}, \tag{7.54}$$

where E_f, E_b, E_p represent the error scores for the main forecast, the benchmark forecast and the perfect forecast respectively. The error measure could be any of those presented above, such as RMSE, or MAE for point forecasts, or CRPS for probabilistic forecasts.

Often the error should be zero for a perfect forecast, and in this case the skill score reduces to

$$SS(\hat{\mathbf{L}}, \mathbf{L}_b) = 1 - \frac{E_f}{E_b}. \tag{7.55}$$

The skill score can obtain a maximum value of 1 if the forecast is perfect ($E_f = 0$), but is equal to zero if is only as good as the benchmark, and of course the score can be negative if the forecast is worse than the benchmark.

The benchmark here is often called a standard, or reference, forecast, but the important point is that this method is kept constant to allow more appropriate comparisons. Since the benchmark methodology stays the same then this allows a comparative analysis of the forecast across different datasets as well as over different

periods in the same dataset. If two datasets have very different "predictabilities" then you can compare the performance on them better via a skill score since the forecast error on the less predictable dataset will be scaled according to the common benchmark, which will also perform more poorly on this data set relatively to the other. As a consequence, the relative performance on the datasets can be more easily compared and will not simply be based on bad luck due to the features of the test dataset that is used.

The main question when creating a skill score then, is what is an appropriate benchmark to use? The following is some suggested criteria:

1. It should be quite simple and not require too much data or additional data sets to produce. This enables the model to be used in most circumstances.
2. It should be easy to implement so that other forecasters can easily replicate it.
3. It should be easy to interpret to help with model evaluation and improvement.
4. It should not be too sophisticated, or state-of-the-art. It is only needed for comparison and hence there is no need for a complicated or "difficult to beat" model.

For many applications, the simple benchmark models described in Chap. 9 will be sufficient. The persistence model is quite common. Other considerations about choosing appropriate benchmarks are given in Sect. 8.1.1.

7.5 Residual Checks and Forecast Corrections

Ideally a forecast is a good estimator of the true load but for various reasons may require some corrections. Common in climate modelling is a model bias where the mean (or expected) value of the prediction is consistently shifted from the actual values. Analysing the residuals of the final forecast model is a common way to both evaluate your model and identify possible improvements for future implementations.

Whatever models are created for time series forecasting there may still be some structure remaining in the residuals which could be exploited to further improve the accuracy of the forecast. Suppose a forecast model is generated for the time series L_t over the time steps in the training data $t = 1, \ldots N$. Let $\hat{L}_t$ represent a forecast estimate fitted to the training data and recall from Sect. 5.2 that the residual time series can be defined as $r_t = L_t - \hat{L}_t$ for $t = 1, \ldots N$. A desirable feature for a forecast model is that this residual series is essentially random noise, since any remaining patterns/relationships could be used to improve the forecast.

The first check should be to plot the residual time series and look for any remaining patterns or features. If the model has correctly explained the data, then the residual series should be random noise,[2] in other words, their values are independent and identically distributed with zero mean.

[2] The noise can follow a particular distribution even if it is random. White noise, is noise which is distributed according to a Gaussian function.

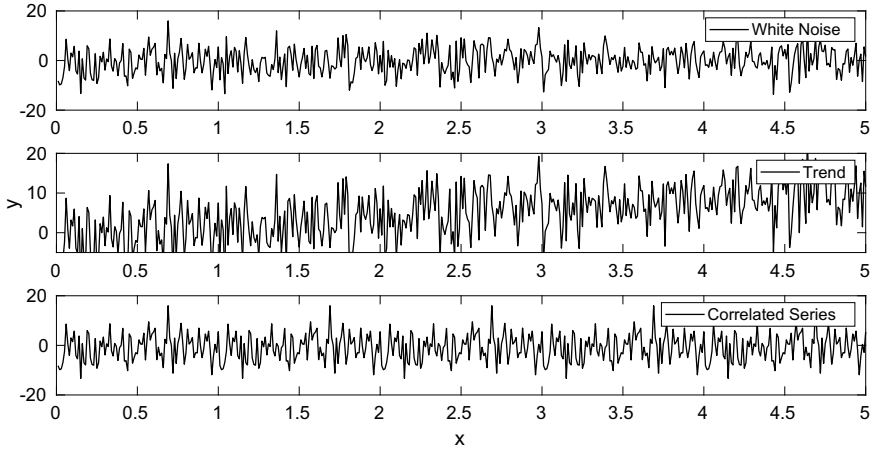

Fig. 7.6 Examples of three time series: white noise (top), white noise plus trend (middle), and a periodic time series based on the first white noise series (bottom)

If two random variables X and Y are independent it means, unsurprisingly, that the value of one is unrelated to the value of the other. In the language of probability this means for all x, y the events $X \leq x$ and $Y \leq y$ are independent. In other words the joint distribution, $F_{X,Y}(x, y)$, of X, Y is related to the individual distributions via

$$F_{X,Y}(x, y) = F_X(x)F_Y(y), \tag{7.56}$$

where F_X and F_Y are the cumulative distribution functions for X and Y respectively (see Sect. 3.3). For an independent variable the correlation between the values is zero. However, note that the reverse is not necessarily true, zero correlation does not imply they are independent. However, it can be used as evidence for independence, or at least increase the plausibility that they are independent.

It is usually quite easy to tell if the data is not white noise but not trivial to test if it is. A plot of the time series should give an initial indication of which case may be true. Examples of a few residual time series are shown in Fig. 7.6. In this example, the values at all time steps have the same variance, hence the only thing to check is whether they have zero mean and are independent.

In this Fig. 7.6 there are two that look like white noise and one that is quite clearly not white noise. The top plot is white noise, the middle plot is the original white noise series but augmented with an increasing trend. The bottom series looks like it is white noise but however is formed by repeating a chunk of 100 data points from the top series. Therefore in fact it has strong autocorrelation despite not being directly obvious.

The autoregressive features in this series can be confirmed by looking at the autocorrelation plot (As when creating the ARIMA model in Sect. 9.4). In this case the periodic time series does not have components which are independent as seen by

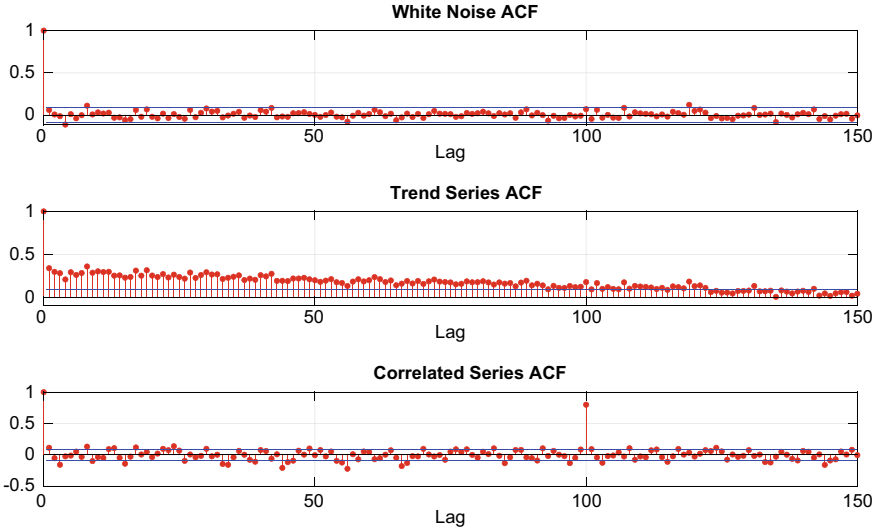

Fig. 7.7 ACF's of the three "white noise" time series: white noise (top), white noise plus trend (middle), and a periodic time series based on the white noise series (bottom)

the spike at lag 100 in Fig. 7.7. The ACF plot shows that the true white noise series has no autocorrelation as expected, but the noise with trend in the middle does show up via the slow decay in the ACF as a function of lag.

Any autocorrelation which remains in the residuals can be removed by including extra autoregressive components to the residual series (alternatively can be added to the original model) via

$$r_k = \sum_{k=1}^{p_{\max}} \phi_k r_{t-k} + \epsilon_t \tag{7.57}$$

for some assumed Gaussian error ϵ_t and optimal autoregressive order p found by minimising the Akaike information criterion (AIC) or Bayesian information criterion (BIC) (see Sect. 8.2.2) over $p \in \{0, \ldots, p_{\max}\}$, for some maximum order $p_{\max}$. After training the coefficients $\phi_1, \ldots, \phi_{\max}$, a forecast for the residual can be produced via $\hat{r}_k = \sum_{k=1}^{p_{\max}} \phi_k r_{t-k}$ and the original forecast can be updated to achieve a new forecast for the load series via $\hat{L}_t + \hat{r}_t$. If sufficient lags have been included in the updated model, the new residual series $\tilde{r}_t = L_t - \hat{L}_t - \hat{r}_t$ should now have no significant autocorrelations. Additionally it is hoped that the new models will have improved forecast accuracy. Of course another autocorrelation check of the residuals can be performed on the new forecast and the process repeated if not all of the autoregressive features have been accounted for.

Another form of bias in a time series forecast is whether the residuals are centred around zero. The random noise with trend is one such example. The simplest form of bias is where the noise is centred around a non-zero constant. This suggests a

simple bias correction can be applied to reduce the mean error to zero by shifting by the sampled mean of the residual. I.e. set $\hat{r}_t = r_t + b$ where $b = -\frac{\sum_{t=1}^{N} r_t}{N} \in \mathbb{R}$. Alternatively one can directly update the forecast itself, i.e.

$$\hat{\hat{L}}_t = \hat{L}_t + b. \tag{7.58}$$

Thus by definition the new residual series has sample mean equal to zero. The residual time series with linear trend can also be corrected in a similar way except by detrending with the line of best fit through the points (similarly as seen for the outlier detection in Sect. 6.1.1). More generally, where the trend is obvious, similar detrending approaches can be applied.

There may also be assumptions concerning the distribution of the residuals for particular models. For example linear regression and ARIMA models assume Gaussian distributions. If the residuals are not distributed symmetrically then a transformations of the data may be required (Sect. 6.1.3). Further non-constant variance of the residuals suggests that methods which assume fixed variance may not be appropriate. Instead, alternative approaches such as the GARCH type models introduced in Sect. 11.6.2 may be required.

In general applying forecast correction and checking for independence is not straight forward. As shown above, time series plots of the residuals should be the first consideration and then checks for constant mean and variances can be performed by calculating them on fixed intervals of the residual time series and comparing them to the full sample mean and variance. Finally, the autocorrelation and partial autocorrelation functions should be plotted to check for moving average and autoregressive components and identify dependence between points in the time series.

The above methods are primarily focused on point estimates. However, for probabilistic forecasts there are also corrections which can be applied, but they are often more complicated than point forecast corrections. For a simple case recall in Sect. 7.2 that a probabilistic forecasts should have a uniform probability integral transform, but if this is not the case, then the PIT can also suggest ways to inform possible corrections. For example, as seen in Sect. 7.2, overdispersed (alternatively underdispersed) forecasts produce a wider (or narrower for underdispersed) PIT distribution than is desired, which means the model could be improved by squashing (or stretching for underdispersed estimates) the distribution. More generally we can look at the PIT to see which areas of the distribution are over or under represented. There are more sophisticated calibration methods such as quantile mapping which have traditionally been applied in climate and weather modelling, further reading in these areas are given in the Appendix D.

7.6 Questions

For the questions which require using real demand data, try using some of the data as listed in Appendix D.4.

1. Take a few days from a real demand time series and create basic forecasts by shifting the profile by full day. Calculate the MAPE, MAE and RMSE. Compare them. Take a hundred smart meter time series and calculate the errors based on the same seasonal persistence forecast model. Produce a scatter plot of the errors against the size of the demand (e.g. the average half hourly or daily demand). Is there a pattern you notice in the plots? If you plot the time series of the profile against the forecasts can you identify the sources of error for those with the best and worst accuracy?
2. Take a half hourly household demand profile with a peak in the evening. Take a day and shift the profile by an hour in one direction (add the shifted points that fall off the end to the other side). Now calculate the RMSE error between them. Next produce a flat profile by taking the average half hourly value and setting all half hours of the day to this value. Calculate the RMSE between this and the original profile. Compare the two error values. Which is smaller? Try this with several other forecasts. Is the flat profile producing smaller errors than the shifted in some cases? This is explored more in Sect. 13.3.
3. Sample 5000 points from a univariate Gaussian distribution. Select quantiles at $0.05, 0.1, 0.15, \ldots, 0.95$ and plot the PIT. How many points should be in each quantile range? Now delete 5–10 points from the middle five quantiles of the distribution. Plot the PIT again, how has the shape changed? Is it underdispersed or overdispersed? Repeat the experiment but remove values from the tails of the distribution. Replot the PIT and check whether the shape is underdispersed or overdispersed. Now plot the reliability diagrams for all three samples (this will require calculating the empirical quantiles for each sample).
4. Sample 1000 points from a univariate distribution of your choice. Create three empirical distributions from these samples by deleting the same number of points (say 10%) (a) randomly, (b) from the centre of the distribution, and (c) from the tails of the sample. Use the samples to calculate quantiles which will now define your probabilistic estimates. Calculate the pinball loss score for these three distributions on the original sample of points. Repeat the calculation for the CRPS. Which has the best (lowest) score?
5. Consider forecast errors with horizon. Take some half hourly or hourly demand data. Create a simple forecast of the next two weeks by repeating the daily profile for one day, for the next fourteen days. Calculate the RMSE error for each day. How does it change with horizon? Repeat this with other time series and observe the change with horizon. Does it change smoothly with how many days ahead? Or is there a change depending on the day of the week?

6. Take the forecast used in the last question. Produce the residual time series. Plot the autocorrelation and partial autocorrelation plots. Which lags produce the biggest coefficient values? How many lags would you therefore expect to need to correct for this in an autoregressive update to this model? If you know how to apply linear regression try adding these terms to your model and repeat the forecast again. How have the errors changed? If you don't know how to apply this, you can wait until you've read Chap. 9 and come back to this part of the question!