TECHNICAL MEMORANDUM

766

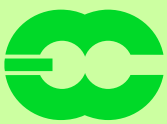
Significance of changes in medium-range forecast scores

Alan Geer

Research Department

Submitted to Tellus

October 2015



European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen terme

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/en/research/publications

Contact: library@ecmwf.int

European Centre for Medium-Range Weather Forecasts
Shinfield Park, Reading, RG2 9AX, England

Abstract

The impact of developments in weather forecasting is measured using forecast verification, but many developments, though useful, have impacts of less than 0.5% (about 0.5 h) on medium-range forecast scores. Chaotic variability in the quality of forecasts makes it hard to achieve statistical significance when comparing these developments to a control. For example, with 60 separate forecasts and a 95% confidence interval, a change in quality of the 5-day forecast would need to be larger than 1% to be statistically significant using a Student's t-test. The first aim of this study is simply to illustrate the importance of significance testing in forecast verification, and to point out the surprisingly large sample sizes that are required to attain significance. Further, chaotic noise is correlated in time and can generate apparently systematic-looking improvements at different forecast ranges, so a 'run' of good scores is not necessarily evidence of statistical significance. Even with significance testing, forecast experiments have sometimes appeared to generate too many strange and unrepeatable results, and a second aim has been to investigate this. By making an independent realisation of the null distribution used in the hypothesis testing, using 1,885 paired forecasts (about 2.5 years of testing), it is possible to construct an alternative significance test that makes no statistical assumptions about the data. This is used to experimentally test the validity of the normal statistical framework for forecast scores, and it shows that the naive application of Student's-T does generate too many false results. A known issue is temporal autocorrelation in forecast scores, which can be corrected by an inflation in the size of the error bars, but typical inflation factors (such as those based on an AR(1) model) are not big enough and are not reliable for smaller samples. Also, the importance of statistical multiplicity has not been appreciated. For example, across three forecast experiments, there is a 1 in 2 chance of getting a false result through multiplicity. The t-test can be reliably used to interpret the significance of changes in forecast scores, but only when correctly adjusted for autocorrelation, and when the effects of multiplicity are properly considered.

1 Introduction

Forecast verification is central to the ongoing development of weather prediction. Every year, a Numerical Weather Prediction (NWP) centre will introduce dozens of scientific developments that need to be tested and then merged into an upgrade of the forecasting system. The quality of medium-range forecasts in the northern extratropics is improving by around 1 day (about 20%) per decade at centres such as the European Centre for Medium-Range Weather Forecasts (ECMWF, e.g. Simmons and Hollingsworth, 2002; Magnusson and Källén, 2013). Although some of that comes from step-changes like the introduction of variational data assimilation (e.g. Andersson et al., 1998) most of the improvement likely comes from the flow of smaller developments like model parametrization upgrades or newly-launched satellites. These changes might have an impact of less than 30 minutes of forecast skill, or about 0.5%, which is hard to detect with forecast verification (note also that skill improvements do not necessarily add linearly, so it takes many smaller developments to achieve a 20% improvement per decade.)

A main issue is that chaotic error growth leads to large differences in skill between any two forecasts (e.g. Hodyss and Majumdar, 2007). Even when verification from hundreds of forecasts is combined these chaotic differences are large enough to cause appreciable differences in average forecast scores. As will be shown, two perturbed but otherwise identical model configurations can differ by 1% (or one hour) in forecast error when computed from a sample of 60 forecasts. A real change in error of 0.5% is undetectable in this context. To verify any scientific change at ECMWF it is now routine to use at least 360 forecasts. One aim of this study is to investigate the size of sample required to reliably test a development in a high-quality forecasting system. A broader hope is to encourage better practice in evaluating scientific developments in forecasting which are sometimes documented without consideration of statistical significance or the limitations imposed by the chaotic variability of the atmosphere.

If statistical significance is sometimes neglected, it may reflect a lack of available documentation. Although a significance test of paired differences is routinely used at forecasting centres, and the fundamentals are covered by Wilks (2006), there is little else in the literature on forecast verification. At ECMWF, the paired difference approach was set in place by Mike Fisher (1998), though in an internal publication. A textbook on forecast verification by Jolliffe and Stephenson (2012) devotes only a few pages to the subject of significance testing. It is hoped that the current work can serve as a reference, so it will summarise the basics (in Sec. 3.)

Even when significance testing is applied, scientists question its reliability: some think the significance tests are over-cautious but others wonder if some unusual statistical property of forecast scores might cause significance to be overestimated. This author's suspicion was that there are too many unrepeatable results. A recent example was when a set of 360-sample experiments needed to be repeated due to a compiler bug, and the statistically significant impact at day 7 moved from the northern hemisphere (NH) to the southern hemisphere (SH), suggesting at minimum that no significance was associated with the hemisphere in which the impact appeared (Geer et al., 2014). Although the compiler bug might have affected forecast quality, it is most likely that the tests of statistical significance were too generous. Hence a second aim is to explore the validity of the current approach to statistical significance testing of changes in forecast error, which has not been fully established. Fisher (1998) showed that temporal autocorrelation was important, and added a correction term to the ECMWF 'Verify' package, but the 'Iver' package has not applied a correction. The current study will also investigate non-Gaussianity and multiple comparisons as possible explanations.

This study focuses on medium-range (day 3 to day 10 here) extratropical forecast scores of which the quintessential example is the root-mean-square (RMS) error in 500 hPa geopotential height (e.g. Lorenz, 1982). These scores are principally sensitive to errors in the location of the low and high pressure features of extratropical Rossby waves. Studies including Murphy (1988) have illustrated some of the pitfalls of using RMS error as a measure of forecast skill. Broadly, RMS error is sensitive to systematic ('unconditional') biases between the forecast and the verifying reference, as well as to 'conditional biases', i.e. differences in the standard deviations of the forecast and the reference field. RMS errors also change according the the 'activity' of the atmosphere, i.e. the standard deviation of anomalies compared to climate mean (e.g. Thorpe et al., 2013). However, because baroclinic errors grow so rapidly, RMS errors are a reliable indicator of forecast quality in the midlatitudes in the medium-range. Systematic errors and errors in the verification reference (whether analysis or observations) are small by comparison. Hence also, for the scores examined here, standard deviation and RMS are near-identical, so the distinction is not important. Conditional and unconditional biases do cause confusion in the interpretation of forecast scores in the early forecast range and wherever systematic errors are comparable to random errors, particularly in the tropics and the stratosphere. For these reasons, this study explicitly avoids the tropics, stratosphere and short-range forecasts (i.e. before day 3). Also, this study does not consider ensembles of forecasts, though clearly they could help increase the sample size and reduce the size of the error bars. At the moment the computational cost of an ensemble forecast system is too high for routine testing.

The error of synoptic forecasts is not the only measure of forecast skill: many users of forecasts are more concerned with variables like near-surface temperature, cloud and rain. However, it is still important to check that upgrades do not cause harm to the large-scale forecast. Problems can come through scientific oversights, unforeseen interactions or simply from coding errors. The intention is not to exclude other valuable approaches in forecast testing and verification, but any improvement in medium-range forecasting relies on a good synoptic forecast, so these basic scores will always be important.

As mentioned, the usual basis of significance testing is to create a population of paired differences in scores between an experiment and a control (e.g. Wilks, 2006). The point of significance testing is to

Table 1: Experiments used in this study

Name	Description
Control	ECMWF operational configuration (cycle 38r2) with all observations but with T511 horizontal resolution.
AMSU-A denial	The AMSU-A on Metop-A has been removed from the observing system.
Noise	Scientifically identical to the control, but with a technical modification that results in tiny numerical differences in some operators in the data assimilation, compared to the control.

reject the 'null hypothesis', i.e. the possibility that a scientific upgrade has no effect on forecast scores. To do this we need to estimate the properties of the 'null population', which in our case is the variations in skill that can be produced by a forecasting system in which nothing has changed scientifically. In a paired difference approach the null population is not available and its variability is estimated from the standard deviation of the population of paired differences. However it is totally feasible, if expensive, to explicitly compute a null population using perturbed but otherwise identical experiments. The main advantage this study has over any previous one is that it is based on a sample of 1,885 forecasts covering two and half years of experimentation. From this it is possible to generate an independent realisation of the null population that can be used to experimentally test the assumptions of Gaussian and uncorrelated errors that are being made in the classical approach. It can also be used to create an entirely non-parametric significance test, i.e. one that makes no assumptions on the distributions of forecast scores. These are the new tools that will be used in this study.

2 Experiments

In the experiments presented here, 1,885 forecasts are available, initialised twice-daily at 00Z and 12Z from assimilation runs covering 1st January 2011 to 31st July 2013. ECMWF cycle 38r2 is used, which was operational during 2013 and uses 137 vertical levels. A T511 horizontal grid equivalent to about 40 km resolution has been used, lower than the operational resolution but only a little less skilful at the medium range. Forecasts are verified at ranges from 12 h through to 10 days, against the experiment's own analysis, so some forecasts at the end of the period cannot be verified. The precise number of forecasts verified varies from 1,885 to 1,866 for the 12-hour to 10-day forecast.

An example of a routine scientific experiment has been created by removing one Advanced Microwave Sounding Unit (AMSU-A, on the satellite Metop-A, Robel, 2009) from the global observing system. AMSU-As measure temperature in broad vertical layers across the vertical extent of the atmosphere and AMSU-As on seven satellites are the most influential group of observations for medium-range forecasting (e.g. Radnoti et al., 2010). But the typical decision being made at an operational centre (or for the future observing system) is not whether this type of observation should exist at all, which would cause a large impact on scores, but whether to add one extra satellite. Borrowing an economics concept, what matters is marginal impact. As will be shown, going from six to seven AMSU-As gives a 0.5% improvement in forecast scores at day 5, which would be a useful contribution to the (roughly) 2% per year improvement in scores from all developments in data assimilation, modelling and the observing system.

Table 1 summarises all three experiments used in this study; the noise experiment will be introduced later.

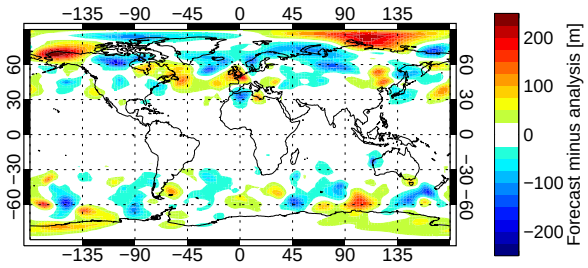


Figure 1: Day-5 forecast minus analysed geopotential height at 500hPa in the control experiment, at 00Z on 31st January 2012.

3 Basics

Forecasts are generated from a set of assimilation experiments and verified against the best estimate of the true atmospheric state, which is usually an analysis. For a field of m model grid-points j, forecasts f_j are compared to analyses a_j valid at the same time. The RMS error r can be computed as:

$$r = \sqrt{\sum_{j=1}^{m} w_j (f_j - a_j)^2} \tag{1}$$

Weights w_j sum to 1 and are proportional to the area of the grid-box. Statistics are usually computed over set regions of the globe, here the northern and southern hemisphere (excluding the tropics) defined as 20°N–90°N and 20°S–90°S. Atmospheric fields are interpolated to a regular latitude-longitude grid, in this case 2.5°, before computing the RMS error. This study concentrates on RMS scores which are the simplest to compute and most popular; other choices such as anomaly correlation or standard deviation can be used, but when considering midlatitude synoptic scores the results are very similar - this will be justified later.

Figure 1 shows forecast minus analysed 500hPa geopotential height on a typical day in the northern winter. These are the $f_j - a_j$ available to Eq. 1, though note the weighting term w_j reduces the influence of the errors at high latitudes compared to their visual prominence in this projection, particularly poleward of 75°. Errors are spatially quite smooth and contiguous across areas at least 1000 km across. It is clear that day 5 forecast errors in 500 hPa height measure inaccuracies in the large-scale synoptic forecast in the midlatitudes. Small-scale errors have little impact on the RMS and there is no need to use a higher resolution than 2.5° to compute these scores. Tropical regions exhibit little variability in geopotential, hence the relative lack of errors there. The situation is similar through the medium-range but in the short-range, errors are generally on scales of 500 km or less (not shown). Interestingly, day-5 wind errors exhibit smaller scales (e.g. 500 km, not shown) but as will be seen, regional wind scores are highly correlated with geopotential height errors.

The RMS error as a function of lead-time (Fig. 2) illustrates the rapid baroclinic error growth in midlatitudes. As mentioned, the experiment's 'own analysis' has been chosen as the verification reference. Where the experiments are comparable in skill and close in quality to the best available, the own analysis is a reasonable choice of reference because it avoids various possibilities of biasing the comparison in the early forecast range (e.g. Geer et al., 2010). However, the forecast error at the initial time is by definition zero, which is unsatisfactory and incorrect. Nevertheless, issues of short-range verification are not the subject of the current study. All results and figures in this study have also been repeated using the ECMWF operational forecast as the reference. However, such a change in the reference has negligible

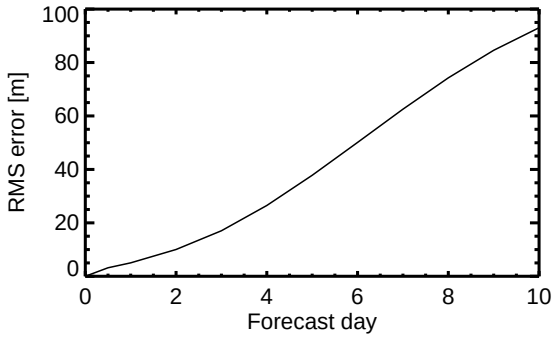


Figure 2: RMS error of NH geopotential height at 500hPa in the control experiment, as a function of forecast lead time, for the control experiment, aggregated over 1st January 2011 to 31st July 2013.

effect on this study. Appendix A gives further details.

A noteworthy aspect of the statistical testing framework is that each forecast is verified individually, and the scores are averaged arithmetically. Figure 2 shows the mean of the RMS over the n forecasts (1,875 at day 5 in the example) with i representing each forecast:

$$\bar{r} = \frac{\sum_{i=1}^{n} r_i}{n} \tag{2}$$

The main reason to treat the r_i as separate samples is to provide a population to which the Student's t-test can be applied. An alternative approach would be to compute the RMS (i.e. Eq. 1) across all grid-points and forecast times simultaneously, but that would make it more difficult to devise a hypothesis test, and the results would likely be dominated by large winter errors (in the same way that an RMS computed globally over Fig. 1 would be dominated by errors of the NH winter.)

Figure 3a shows the evolution of day-5 RMS error of northern hemisphere (NH, 20°N - 90°N) geopotential height in the control run. The most obvious feature in RMS error is the seasonal cycle, with errors reaching 65 m in winter and sometimes reducing to 20 m in summer. This reflects the larger amplitude of Rossby waves in winter. It is clear that forecast scores tend to be correlated from one day to the next, with periods of smaller or larger errors persisting for at least a week, reflecting the amplitude and predictability of the evolving weather patterns. Autocorrelation between RMS scores of successive forecasts (every 12 h) and every second forecast (every 24 h) are 0.76 and 0.6 respectively. For this reason, scores are highly correlated between different experiments, as illustrated in panel b. The scientific components in these experiments (e.g. forecast model, the vast majority of observations) are mostly identical and that must also contribute to the correlation between them. Still, the denial of an AMSU-A is a fairly important scientific perturbation of the system but it seems to make little difference to the forecast scores. There is little hope of applying a t-test to these results: control and experiment are highly correlated, and (not shown) their PDFs are not Gaussian, having a fat tail in the direction of higher errors, i.e. towards forecast 'busts'. This reflects the non-linear, chaotic nature of synoptic forecasts.

The solution (Fisher, 1998; Wilks, 2006) is to look at the paired differences, i.e. the time-series of:

$$d_i = r_i^{\mathrm{EXP}} - r_i^{\mathrm{CONTROL}} \tag{3}$$

This is shown in Fig. 3c; seasonal variability and forecast-to-forecast correlation appear to have been removed. The time-autocorrelation between paired differences at day-5 is 0.15 for base times separated by 12 h and 0.07 for separations of 24 h. Autocorrelation is essentially zero for any longer period. Figure 4

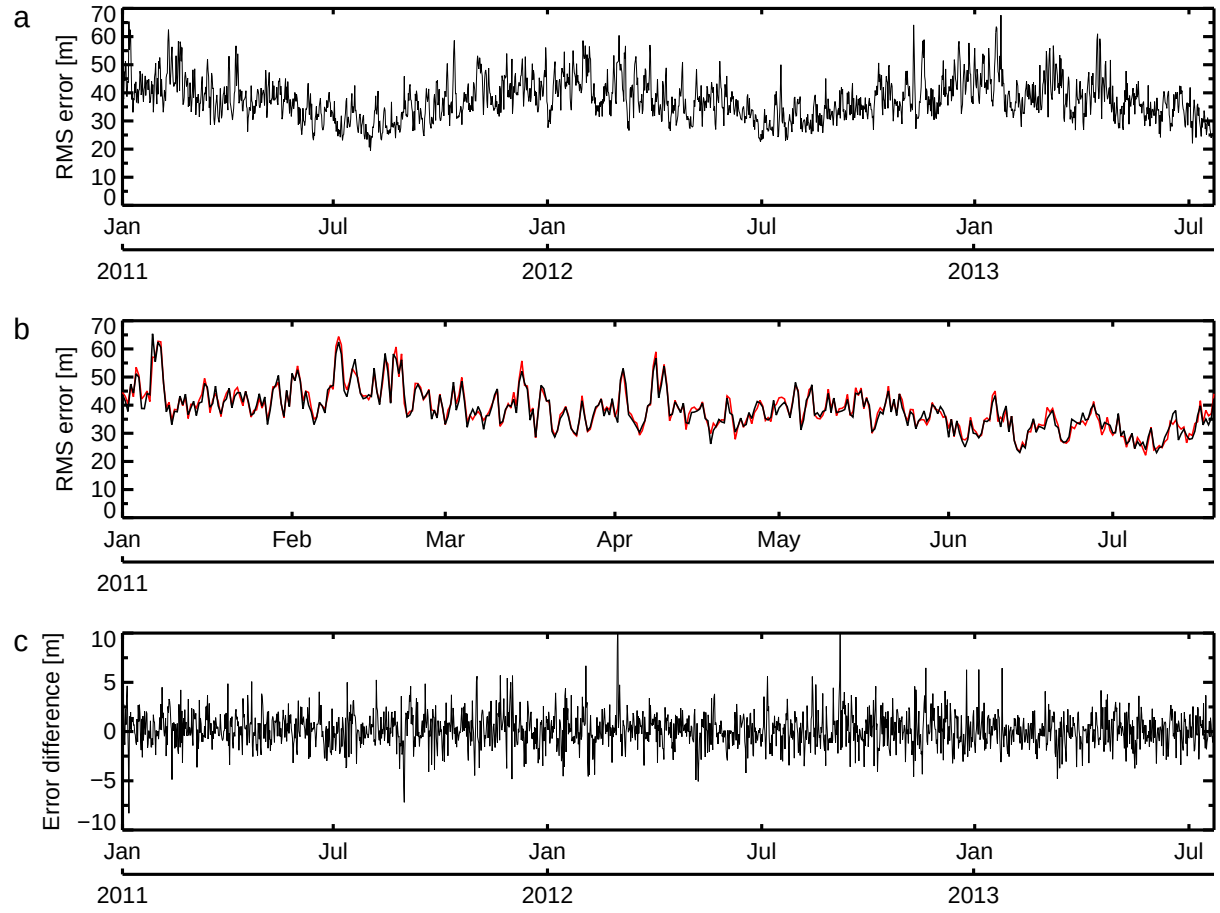


Figure 3: Timeseries of (a) RMS error of 500 hPa NH geopotential height at day 5 in the control experiment (black), verified against an own-analysis reference; (b) As (a), but covering a shorter period of time and including scores from the AMSU-A experiment (red); (c) Experiment minus control over the full time period.

summarises the autocorrelations across different forecast validity times. As shown by Fisher and later in this study, to get the statistical testing exactly right, it is important to deal with the autocorrelation. However, for now we will assume that taking paired differences has created a statistical population that has sufficiently little autocorrelation that a normal t-test can be used to test for statistical significance.

We want to know how the forecast scores of the experiment differ from the control and the usual way is to compute the mean of the paired differences across the n forecasts:

$$\overline{d} = \frac{\sum_{i=1}^{n} d_i}{n} \tag{4}$$

In this example $\overline{d}$=+0.18 m, so eliminating one AMSU-A appears to degrade the scores, i.e. it makes RMS errors slightly larger. However, this is small in comparison to the standard deviation s of the paired differences, which is 1.76 m in this example. The question is whether this 0.18 m increase in error is statistically significant.

We can set up a null hypothesis, which is that forecast-to-forecast variability in d_i is purely random and centred on zero. More simply, the null hypothesis is there is no mean difference in scores and that any variability in scores comes from the natural chaotic variability of forecast quality. The null hypothesis gives rise to a null population, which is an infinite set of random differences between forecast scores d_i.

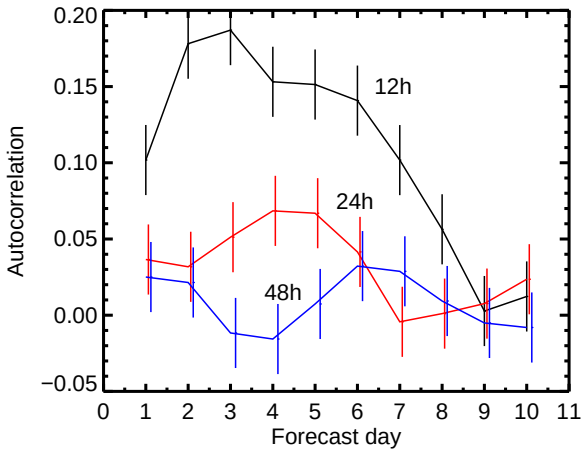


Figure 4: Time autocorrelation of paired differences (AMSU-A Denial minus Control) in RMS error of NH geopotential height at 500hPa, as a function of forecast lead time. Autocorrelations are shown for forecasts separated by 12 h (black), 24 h (red) and 48 h (blue). This pattern is representative of all tropospheric levels in the NH and for other choices of forecast score, such as anomaly correlation, or the RMS error in temperature or vector wind. However, autocorrelation becomes very significant in the stratosphere, where systematic errors dominate RMS forecast scores, and it is slightly larger in the SH. Error bars have been estimated with a Monte-Carlo approach: they are represented by the standard deviation of a large number of lag-1 autocorrelations estimated from samples of 1,885 normally distributed random numbers.

Given the mean $\overline{d}$ and standard deviation s of the paired differences, a test statistic z can be computed (e.g. Wilks, 2006):

$$z = \frac{\overline{d} - \mu_0}{k(s/\sqrt{n})} \tag{5}$$

If the null hypothesis is true, z can be shown to follow a Student's t distribution, which varies according to the population size n but for larger n it tends towards a standard Gaussian (i.e. one with unit standard deviation and zero mean). In a t-test it is necessary to specify the mean of the null population μ_0 but in the current case it is zero and will not be written explicitly from now on. k is an inflation factor to account for autocorrelation or other effects that might affect the results. This will be useful later but for the moment we assume no inflation so $k = 1$. The t-test recognises that a mean computed from a finite sample is unreliable, compared to the true mean of a population. z is essentially the ratio of the mean of the population (e.g. $\overline{d}$) to an estimate of the sampling error in that mean, i.e. $(s/\sqrt{n})$. For smaller n (e.g. less than 50) the t-distribution is broadened compared to a standard Gaussian because s is only a relatively poor estimate of the population standard deviation. However, for the large sample sizes needed for forecast verification, the t-distribution is nearly indistinguishable from a Gaussian.

If z falls outside the typical range of the t-distribution, that shows that the mean of our population $\overline{d}$ is significantly different from the null population. For large sample sizes, 95% of the t-distribution falls within ± 1.96. In our example, $z = 4.52$, which is far outside the usual range of the t-distribution. Hence even though eliminating one AMSU-A increases forecast errors by only 0.18 m, this is still a highly significant result in the context of two-and-a-half years of experimentation.

The difference in forecast scores $\overline{d}$ is usually presented against forecast lead-time (Fig. 5) by normalising with the forecast score of the control, i.e. $\frac{\overline{d}}{\overline{r}_{\mathrm{CONTROL}}}$. Statistical significance is represented with an error bar. The range of t-test values enclosing 95% of the t-distribution can be labelled $\pm z_{95}$ (for larger samples, this is ± 1.96). This is our 95% confidence range in z-space. The confidence range in terms of $\overline{d}$ is

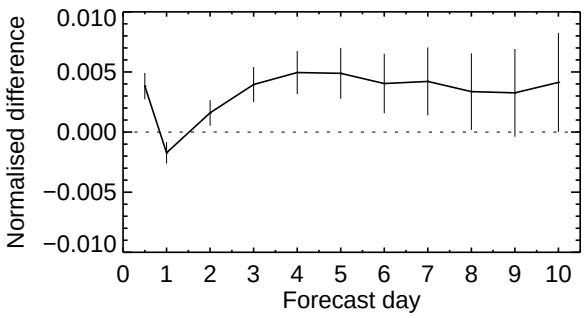


Figure 5: Normalised change in RMS error of NH geopotential height at 500 hPa (AMSU-A denial minus control) as a function of forecast lead time, aggregated over 1,866–1,885 forecasts between 1st January 2011 and 31st July 2013. Error bars give the 95% confidence range.

the maximum value that would be considered statistically insignificant, d_{95} which can be obtained by substitution into Eq. 5:

$$\pm d_{95} = \pm z_{95}\, k \left(s/\sqrt{n} \right) \tag{6}$$

This illustrates how k gives a way of inflating the error bars if needed, but it equals 1 for the moment, indicating we are using a standard t-test. The range $\pm d_{95}$ can be normalised by $\bar{r}_{\mathrm{CONTROL}}$ and centred on the forecast score to produce the error bars plotted in Fig. 5.

Failing to reject the null hypothesis when it should have been rejected (in other words, identifying a random fluctuation mistakenly as a significant result) is known as a Type-I error. With a 95% confidence range we are accepting a 5% chance of coming to this type of wrong result from our experiment (the other type of wrong result, type-II, is introduced later). It is worth noting that this is a 'two-tailed' test, i.e. that we are interested both in situations where the forecast scores get worse (in this example, indicating the AMSU-A that was removed was doing a useful job) or where they get better (here, indicating that AMSU-A was causing problems and should not have been assimilated).

Fig. 5 shows the forecast impact of removing one AMSU-A from the global observing system. Though impossible to identify in Fig. 3, this has a statistically significant impact of around 0.5% increase in forecast error from day 3 through to day 8. But the random variability (s) of forecasts beyond day 8 is so large that even with the big sample used here it is impossible to confirm whether eliminating one of seven AMSU-As affects the scores. There is also some strange behaviour up to day 2, which relates to previously-discussed issues in short-range verification including systematic error, but is not the subject of this study. In most routine verification, it is impossible to achieve a sample size remotely as large as used in this example. For verification over a couple of months with $n = 120$, the error bars would be $\sqrt{1866}/\sqrt{120} \simeq 4$ times larger, as inferred from Eq. 6. There would be no statistical significance in any of the results in Fig. 5. The issues of interest in this study are first, to see how far these error bars can be trusted and second, to demonstrate the size of sample required to generate significant results.

4 The null distribution

It has been explained that the t-test, as used in forecast verification, takes the standard deviation of the paired differences and uses it to estimate the statistical properties of what is assumed to be a Gaussian null distribution. However, it is possible to generate a real null population by making an 'upgrade' to the

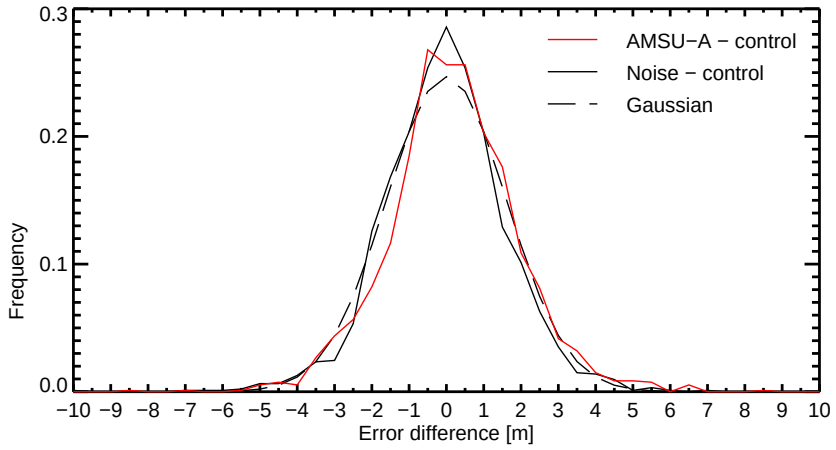


Figure 6: Histograms of paired differences in NH RMS 500 hPa geopotential height errors at day 5. A Gaussian is plotted that has the same standard deviation as the noise minus control population. The integral of this histogram is normalised to 1 (it is a PDF). Bins are 0.5 m in size. Each population has 1,876 members.

forecasting system that has no scientific impact. This will be called the 'Noise experiment' and it can be achieved by making a purely technical change that alters the numerical reproducibility of the forecasting system without altering any of the science[1]. Normal chaotic error growth causes a rapid divergence between any two assimilation experiments into which even the slightest numerical perturbation is introduced – just as seen by Lorenz (1963). Although the observations are the same in both experiments and they keep the analyses on a similar path, the analyses can still take a range of plausible solutions within the range of error of the observations and the model first guess. The forecasts made from those analyses will also vary. The control and the perturbation experiment could be seen as a crude two-member ensemble of data assimilations (EDA). An EDA is used operationally at ECMWF to estimate flow-dependent background errors for the main analysis, and the additional ingredients are broadly just more members, plus perturbations to the observations and to a few model parameters (Isaksen et al., 2010; Bonavita et al., 2015). Due to the additional perturbations applied, the spread of a true EDA is likely to be larger than that between the Noise and Control experiment, so it would not be an appropriate representation of the null distribution we are interested in here.

Figure 6 shows histograms of the paired differences between the AMSU-A experiment and the control (as shown in Fig. 3c) and between the Noise experiment and the control. The histograms are both very similar, and close to a Gaussian. They are not the smooth curves we might hope for, because even with 1,876 samples there are sampling errors in the PDF, compared to one generated from an infinite population. However, the greater errors in the AMSU-A denial experiment show themselves as a general shift to the right compared to the null population. Forecast degradations of greater than 6 m occur on 8 occasions in the AMSU-A denial experiment and on just 1 occasion in the noise experiment. These might be cases where the AMSU-A observations that were removed were particularly critical to

[1]To expand on how the noise experiment has been created: In the ECMWF assimilation system, two otherwise identical experiments exhibit small differences when the number of parallel processes is changed. This comes because a few floating-point summations in the adjoint of the observation operator are made in a different order, resulting in very slight numerical differences. But these small differences grow during the analysis process, resulting in visible differences between analyses. There is a system option that recodes the same sums in a way that can preserve numerical reproducibility, although at additional computational expense, so it is only used for numerical reproducibility testing. However, because we can turn this numerical difference on and off, we can be quite confident that it is fully understood and causes no knock-on bugs or other unknown scientific effects.

Table 2: Statistical properties of the paired differences in NH RMS 500 hPa geopotential height errors at various forecast ranges.

Name	Day	Mean [m]	Std.dev. [m]	Skewness	Kurtosis
AMSU-A - control	3	0.07	0.55	-0.10	2.67
	5	0.18	1.77	0.47	3.60
	7	0.26	3.90	0.21	1.92
	10	0.38	8.43	-0.06	1.27
Noise - control	3	-0.01	0.48	-0.13	1.76
	5	-0.04	1.62	-0.12	1.61
	7	0.12	3.63	0.14	1.10
	10	0.35	8.30	0.08	0.51

forecast skill, perhaps because they observed an area associated with rapidly growing forecast errors. However, it is hard to ascribe statistical significance to that, and mainly, it appears that removing one AMSU-A increases forecast errors by a small amount on average. Apart from the rightwards shift, the PDF of AMSU-A differences looks very similar to that from the noise experiment. Hence, most of the variability in the AMSU-A differences (e.g. Fig. 3c) must come from the same source as that in the noise experiment: random chaotic variability in the quality of forecast scores. In the t-test, we are assuming exactly this: that paired differences in the experiment are composed of a random component that would be the same in the null population, plus a mean shift, which is the signal we are looking for.

Table 2 summarises statistical properties of the populations in Fig. 6, as well as those at other forecast times. As is indicated by the error bars in Fig. 5, at longer ranges it is harder to be sure that the AMSU-A experiment has genuinely larger forecast errors than the noise experiment, because the chaotic variability of scores is so much larger. To underline this, at day 10, the AMSU-A denial and the Noise experiment both apparently cause an increase in error, of very similar magnitude (0.38 m or 0.35 m). The standard deviation of the paired differences for the AMSU-A experiment are similar to those from the Noise experiment, so they appear to be a reasonable estimate of the properties of the null population.

It is worth checking the Gaussianity more rigorously, because non-Gaussian aspects such as 'fat tails' might cause an excessive number of false results in the t-test. The skewness and kurtosis in Tab. 2 indicate minor deviations from Gaussianity, though it appears that any slight non-Gaussianity becomes less of an issue in the longer forecast ranges. Bigger samples should give more reliable results, and this can be achieved by combining populations at different forecast ranges and also from the southern and northern hemisphere, noting that the shapes of the PDFs are similar at different time ranges and in the different hemispheres (not shown). However these scores must not be correlated, otherwise the combined PDF will be artificially broadened (see Wilks, 2006). Section 6 will show that paired differences in forecast scores are correlated to about 5 days in forecast lead-time and across most of the vertical extent of the troposphere. Having excluded the tropics, the stratosphere and the early forecast range (up to day 3 here, for safety) there are only around four independent sets of hemispheric scores available to make a bigger sample: day 4 and day 10 in the SH and NH. These sets are standardised by dividing by their standard deviation and Fig. 7 shows the results. Both the AMSU-A and the noise differences look fairly Gaussian, with skewness and kurtosis relatively close to those expected from a true Gaussian (Tab. 3). A chi-squared test (see Wilks, 2006) would reject the hypothesis that these curves are Gaussian. However, it will be shown later that this level of non-Gaussianity has no real effect on the significance testing of forecast scores, presumably due to the central limit theorem.

It is worth restating the most important conclusion from this section: the differences in forecast scores

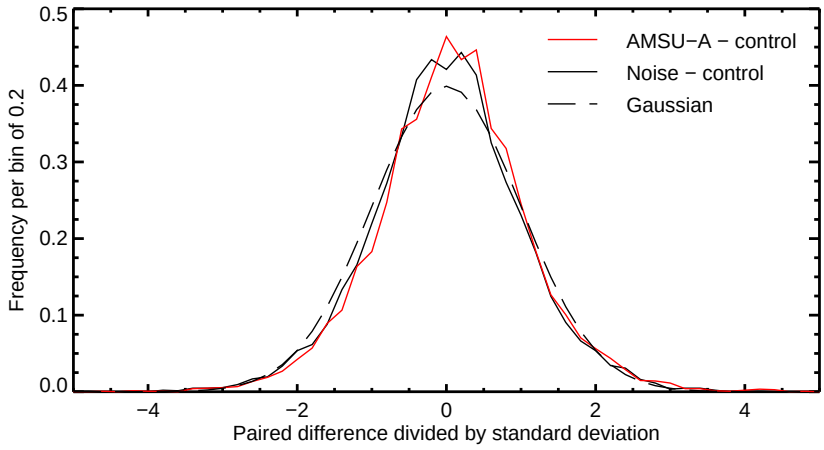


Figure 7: Histograms of paired differences in RMS 500 hPa geopotential height errors, standardised by dividing by the standard deviation of four sub-populations (day 4 and day 10 in both SH and NH). This gives a population of 7,464 in each histogram. The dashed line is the standard Gaussian. Bin size is 0.2 and the integral of the PDF across all bins is 1.

Table 3: Skewness, kurtosis and chi-squared of the populations shown in Fig. 7. A Gaussian will have a skewness of 0 and kurtosis of 0 by the definitions used here. Chi-squared is computed against the Gaussian on the 47 bins between -4.6 and +4.6, and would have to be less than 60.4 to confirm Gaussianity at the 95% confidence level.

Name	Skewness	Kurtosis	Chi-squared
AMSU-A - control	0.145	1.39	1045
Noise - control	-0.053	1.30	168.9

between any pair of experiments show a lot of variability, but almost all of it is down to chaotic variations in the quality of forecast scores. The systematic changes in forecast quality resulting from a typical upgrade to a high-quality forecasting system are many times smaller than the standard deviation of this noise. Identifying small genuine changes in forecast scores is difficult.

5 Establishing significance in shorter experiments

So far we have examined the results from two-and-a-half years of forecast verification, which is unaffordable on a routine basis. What if only 1 month of experimentation, i.e. 60 forecasts, were available? There are 31 different non-overlapping, contiguous blocks of this length in the long experiments, so it is possible to compute the mean change in forecast scores for each of them. Figure 8 shows a histogram of the results from each block, at day 5 in the NH as usual. Over the complete period with 1,876 forecasts, removing an AMSU-A has been shown to increase forecast errors by 0.18±0.04 m. The noise run decreases errors by 0.04±0.04 m. However, removing one AMSU-A would have made no difference, or would have slightly improved scores, in 9 of the 31 blocks. In other words, there is at least a one in three chance of concluding that AMSU-A does not have a positive effect on scores (and the chances are even larger once statistical significance testing is done.) Further, the paired differences from the noise experiment can be anywhere from -0.4 m. to +0.7 m. Had we been unlucky enough to pick the block with a mean of +0.7 m, and had we forgotten to perform any significance testing, we would have concluded that changing the number of parallel processors used in the ECMWF system degrades NH scores by 0.7 m -

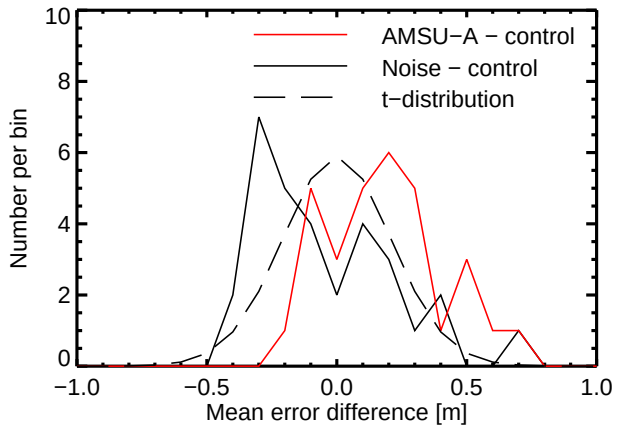


Figure 8: The unreliability of forecast differences computed over one month (i.e. 60 forecasts). The figure shows histograms of the mean of paired differences in NH RMS 500 hPa geopotential height errors at day 5. There are 31 different contiguous 60-forecast samples available in the 2.5 year experimental period. Bin size is 0.1 m. The dashed line shows the t-distribution transformed into the space in which we are computing means.

around 2% or two hours, which is unusually large compared to typical developments. That is nonsense, but it illustrates the dangers in using short periods for verification.

The t-distribution is also shown on Fig. 8, mapped into a distribution of means $\overline{d}$ using using Eq. 5 (using a sample size of 60, a t-distribution with 59 degrees of freedom, and the standard deviation of the simulated null population, which is 1.6 m from Tab. 2.) In broad terms, the distribution of means computed from monthly blocks from the noise experiment is similar to that predicted by the t-distribution, though with only 31 samples available, a good match can hardly be expected. However, the 95% significance level would be similar either with the t-distribution or with the explicitly computed but lumpy equivalent: in either case, it would be around a 0.5 m change in forecast error. Only 2 of 31 monthly blocks from the AMSU-A experiment show an impact larger than this, but so does one block from the noise experiment. In other words there is negligible chance of seeing statistically significant changes in scores from such a small sample of forecasts.

If the t-test is valid for forecast scores, then the block-means computed from the explicit null population should exactly follow a t-distribution. To examine this better, the number of available blocks can be quadrupled by including scores from the NH and SH, day-4 and day-10 as in Sec. 4. Taking samples of 60 contiguous and non-overlapping forecasts, there are 124 different independent t-tests that can be performed on the explicit null population. The z-scores from each of these tests are summarised in Fig. 9a. Making allowances for what is still a very small sample, they are distributed according to the t-distribution. However, there are too many z-values in the tails, i.e. with magnitudes greater than 2. The t-test, as it has been applied so far without accounting for autocorrelation, is too generous and gives more false positives than expected. The cumulative distribution function (CDF) of these z-scores is shown in Fig. 9b and could be used to construct a significance test free from any of the statistical assumptions needed for the Student's t approach. The 95% confidence limits would be at about ± 2.5 rather than ± 2.0 in the t-test. Accounting for this using the inflation factor k in Eq. 6, $k = 2.5/2.0 = 1.25$. Assuming that this result is valid for all sample sizes, the forecast scores presented in Fig. 5 could be revised by boosting the length of the error bars by 25%. This would not fundamentally change the conclusion that denying an AMSU-A causes a deterioration in forecast quality at days 3–7, based on two-and-a-half years of testing. However, the exact length of the error bars can be very important when sample sizes are smaller.

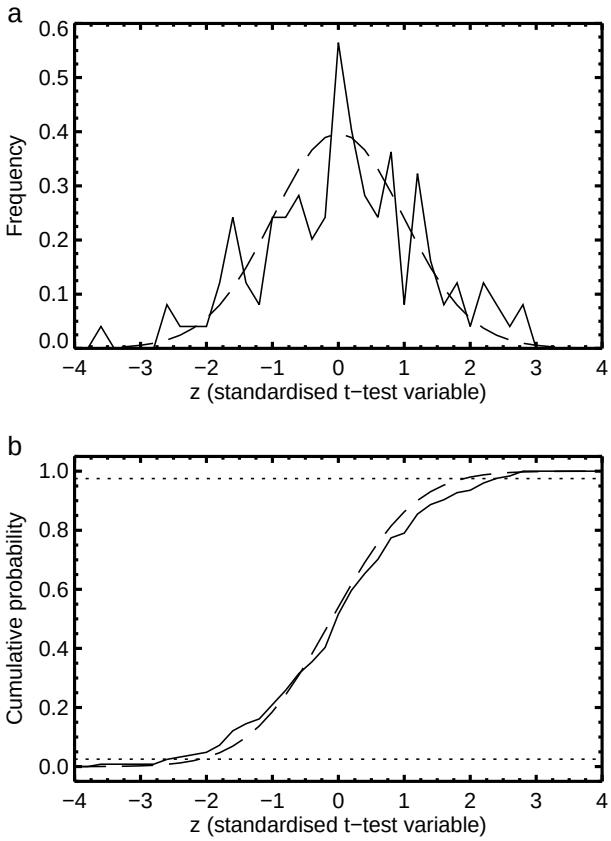


Figure 9: The result of performing t-tests on samples from the explicitly-generated null population, i.e. the differences between the Noise and Control experiments: (a) PDF of z-tests generated from blocks of 60 paired forecast differences (solid) and the Student's t-distribution (dashed); (b) CDF of the same. Dotted horizontal lines indicates the 2.5% and 97.5% quantiles, the limits of the 95% confidence test. Based on a total of 124 independent blocks, 31 from each of the SH and NH at day 4 and day 10. Bin size is 0.25.

The need for an inflation factor due to temporal autocorrelation, and the best way to compute it, will be investigated further in Sec. 7. For now we will concentrate on the even more important issue, which is how hard it is to distinguish the AMSU-A denial experiment from the Noise experiment, in other words, the importance of chaotic variability in the scores computed from 'smaller' samples.

Figure 10 illustrates the improvement in statistical significance when larger samples are used. Here the sample size is 230 forecasts, which is nearly 4 months of experimentation, and divides the 2.5 year experiment period into 8 separate non-overlapping, contiguous blocks. The 95% confidence level from the t-test would be at around 0.21 m. Inflated by 25% based on earlier results, the confidence level would be around 0.26 m. In either case, three of the 8 sections from the AMSU-A experiments would have shown statistically significant results. This is better than with a sample size of 60 forecasts, but still poor odds - we would need an even larger sample to be sure of making the right decision, i.e. that AMSU-A is a useful instrument. Failing to conclude that AMSU-A is useful to forecast scores is an example of Type II error (e.g. Wilks, 2006). The risk of a type II error is hard to estimate a-priori; we are in an unusual situation in being able to estimate it here. For the denial of a single AMSU-A instrument over nearly 4 months of testing, the risk of coming to the wrong conclusion via a Type-II error is 60%, which is still very high.

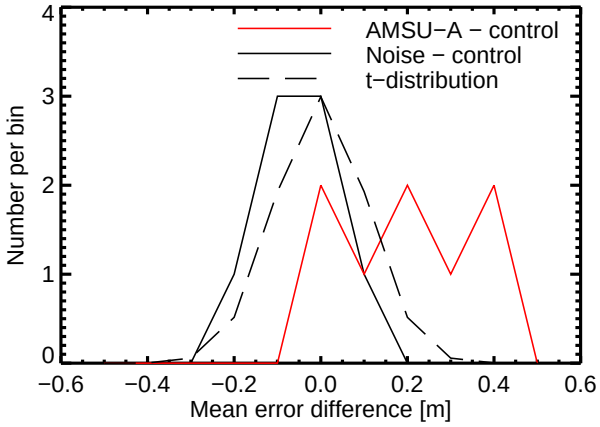


Figure 10: As Fig. 8 but using contiguous blocks of 230 forecasts, giving 8 separate periods for which a mean change in forecast error can be computed. Bin size is 0.1 m.

It is more typical to present normalised changes in forecast scores against forecast range, as shown in Fig. 11a for the 230-forecast samples. Two confidence ranges are included, one based on a t-test without inflation and a broader one with an appropriate inflation factor, based on the null population. For the AMSU-A minus control experiments in panel a, only four of these samples give statistically significant results at day 3 or beyond according to the inflated t-test, giving a 50% risk of a type-II error, i.e. coming to the conclusion that removing an AMSU-A instrument makes no difference to forecast scores. Unlike with the 60-forecast samples, there is little risk of coming to entirely the wrong conclusion that removing an AMSU-A actively improves the scores (i.e. reduces the forecast error). But even a 230-day experiment has too little statistical significance to give reliable conclusions on the impact of an AMSU-A denial.

Fig. 11b shows the change in forecast scores between the Noise and control experiments. The noise experiment should not be mistaken for a genuine change in forecast scores, so it is strange that one of the 8 experiments does show a significant change, according to the inflated confidence range, at day 10. This is an example of a false result, i.e. a type-I error. We were expecting only a 1 in 20 chance of this, given the use of a 95% confidence range. So 1 in 8 seems too high a risk of a false result, though the sample is too small to assess this further. Moreover, to understand this properly it is necessary to bring up the issue of multiple comparisons (multiplicity) which will be covered in Sec. 8. It is also interesting to note that if the t-test were done without inflation (the narrower confidence range), 5 of the 8 noise experiments would show at least one instance of 'statistically significant' improvement. This illustrates how important it is to get the confidence range exactly right, and why narrowly-significant results should be treated with caution. More generally, comparing both panels of the Fig. 11, it would be hard to tell the difference in results between many of the AMSU-A denial experiments and many of the noise experiments.

6 Correlations in forecast scores

So far the focus has been on RMS errors in 500hPa geopotential height. It is also common to verify other dynamical fields (e.g. wind, temperature and relative humidity) and other vertical levels, from the surface to the stratosphere. Another popular measure of error is the anomaly correlation (AC or ACC, e.g. Jolliffe and Stephenson, 2012). Figure 12 correlates the paired differences (d_i, Eq. 3) from these scores against the day 5 RMS errors in 500hPa geopotential height. These results are for the noise distribution (Noise

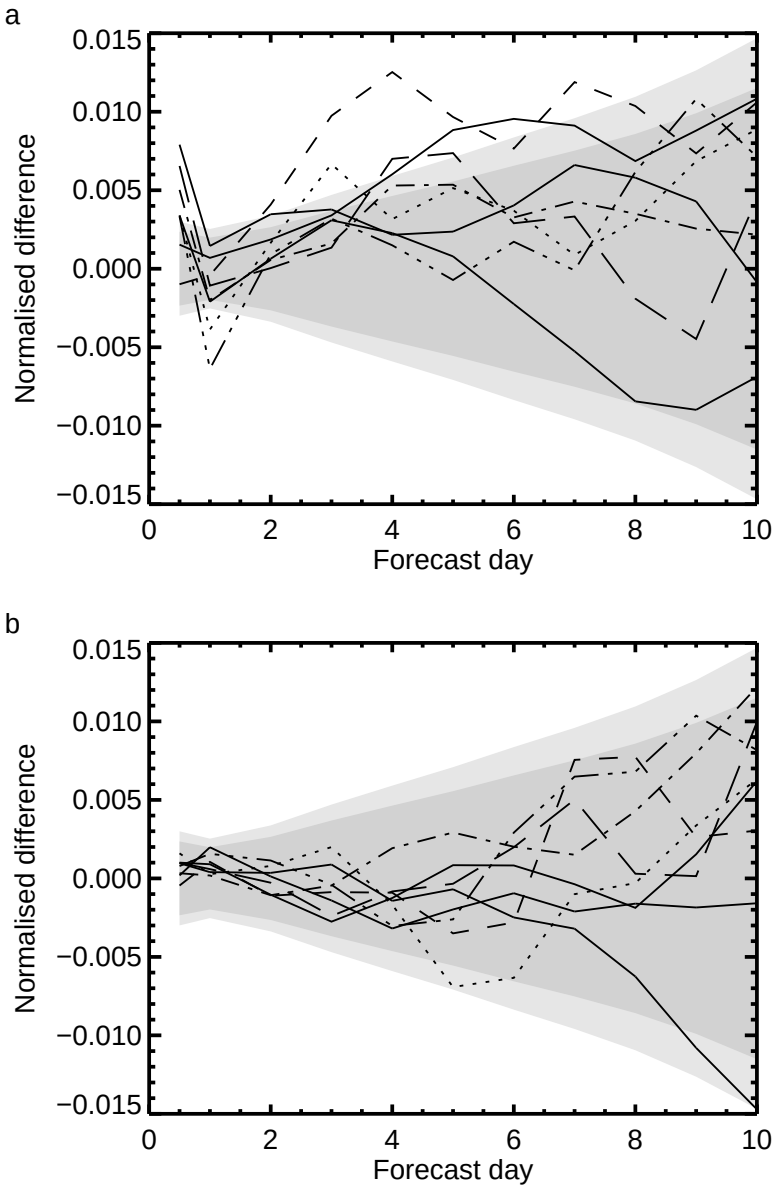


Figure 11: Normalised change in NH 500hPa height RMS forecast error: a) AMSU-A minus Control; b) Noise minus Control. Separate lines are shown for the 8 samples of 230 forecasts. Change in error is normalised by the RMS error of the control experiment. The shading indicates the 95% confidence range estimated in the normal way (no inflation, dark grey) or with the 25% inflation factor estimated from the explicit null population (light grey). To compute the confidence range, the standard deviation of paired differences computed from all 1866 samples in the null distribution has been used. It is centred on zero, which is as valid as centering on the experimental result (the method used in Fig. 5). This approach gives less visual clutter when many experiments are presented on the same figure.

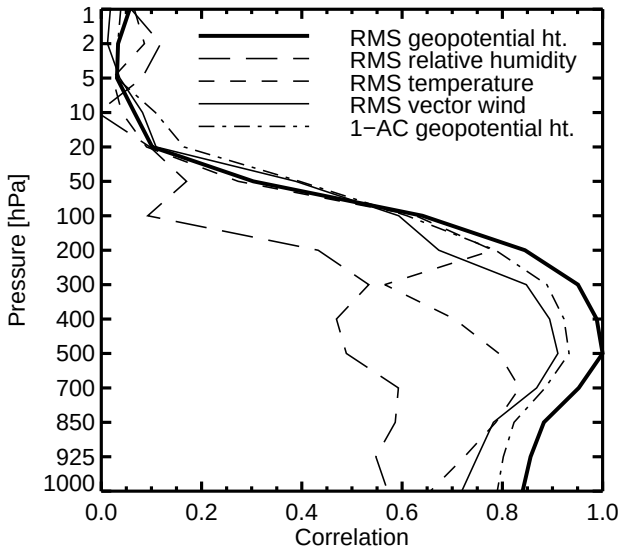


Figure 12: Correlation of paired differences in NH day-5 RMS 500 hPa geopotential height error with paired differences in other measures of forecast error on the same forecast day for the same hemisphere. Anomaly correlation of geopotential height error (AC) has been transformed to (1-AC) to give positive rather than negative correlation.

minus control) but they would be the same for the AMSU-A denial (not shown). RMS geopotential scores are well-correlated across the troposphere though they decline towards the boundary layer and into the lower stratosphere. Anomaly correlation of geopotential height and RMS vector wind errors seem to bring little extra information - they are likely to be measuring rather similar kinds of errors in the large-scale synoptic patterns of the atmosphere (remember that the tropics are excluded from this study). Only the RMS errors in relative humidity seem to be measuring something different, with around a 0.5 correlation to RMS errors in geopotential. A broad conclusion is that results based on geopotential height are likely to be replicated in any midlatitude medium-range score focused on tropospheric dynamics.

Forecast error differences are also correlated across forecast day. If one forecast is better than another at day 5, it is natural to expect the advantage to continue at day 6. That appears to be the case, but correlations across longer time ranges become quite weak. The correlations for RMS 500 hPa geopotential height error are shown in Fig. 13. The correlations of day-5 error differences (the solid line) drop to 0.1 or less at day 1 and day 10. Ignoring correlations less than 0.1, forecast error differences are correlated over about 2 days in the early forecast range and for 4-5 days in the medium range. Hence, even if one forecast is better than another at day 1, as measured by RMS error difference, this says nothing about the RMS differences in the medium range. A hypothesis to explain this is that the large-scale synoptic errors that dominate the RMS scores in the medium or longer range have grown from much smaller, localised errors that are not an important contribution to the RMS scores at earlier ranges. Further, as has been shown, the daily variability of the RMS differences is dominated by chaotic variability.

It is also interesting to look at the consistency of the averaged change in scores ($\overline{d}$, Eq. 4) across forecast time. This analysis has to be performed on blocks of forecast scores like those in Fig. 11. Surprisingly, for blocks containing e.g. 60 or 230 samples, the correlations look a lot like Fig. 13 whether computed from the AMSU-A denial or the Noise experiment (not shown). It should not be misunderstood from these results that initial condition error is not important in controlling the medium range error. Even small changes in the size of the initial condition error do affect the medium range, as can be shown when

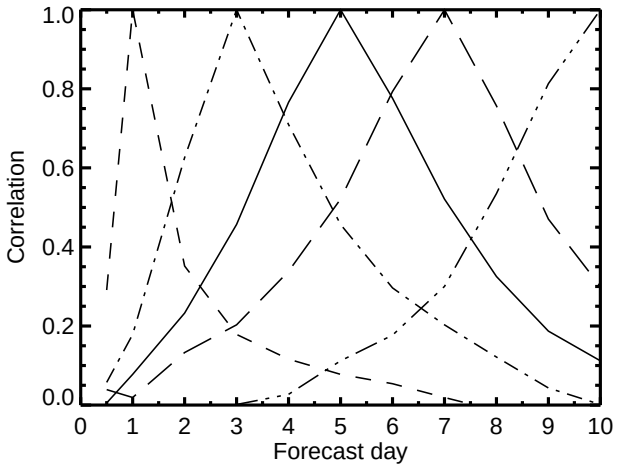


Figure 13: Correlation of paired differences in NH RMS 500 hPa geopotential height error across forecast day, for forecasts with the same base time. Examples shown for D+1, D+3, D+5, D+7 and D+10.

statistics are computed over long periods as exemplified by the AMSU-A denial in Fig. 5. However, the day-to-day variability in paired differences in forecast scores is dominated by noise in most typical experiments.

The effect of chaotic noise on forecast score correlations is best seen in Fig. 11. Panel a shows blocks from the AMSU-A denial experiment. Some blocks show a consistent increase in RMS error across all forecast times, but the smaller details vary in an uncorrelated way. In panel b, the Noise experiment does not show so much consistency across forecast time. However, results can still be consistent across 3–4 forecast days. It is tempting to think that if a few forecast days in a row show consistent changes in the scores, even if these changes are not statistically significant, that makes the results somehow more reliable. Such runs of consistent but not significant changes in scores can be generated by random forecast noise.

Another aim of this correlation analysis has been to find out roughly how many independent statistical tests are being made when we verify a full suite of scores across different hemispheres, different levels and different parameters. This will be particularly important later when trying to understand the impact of statistical multiplicity. Ignoring correlations smaller than 0.1, and excluding the tropics, stratosphere and early forecast range, there are only two completely independent tropospheric scores in one hemisphere: mid-range (day 3-6) and long-range (day 7-10). There is no correlation between paired differences in the SH and NH (not shown) giving a total of four independent scores.

7 The inflation factor

Section 5 has illustrated the need for an inflation factor when using the t-test. Figure 9 suggested that an inflation of around 25% was required, using the explicitly-generated null population and making no statistical assumptions. The assumptions made in the t-test are Gaussianity and (up until now in this study) that there is no temporal autocorrelation in the paired differences between experiments. Fisher (1998) showed the importance of autocorrelation and demonstrated how to compute an t-test inflation factor k using an autoregression AR(1) model. In this method, the correlation between neighbouring

paired-differences is estimated from the population being tested. At ECMWF, this AR(1) inflation factor has been used in the 'Verify' package but not in 'Iver'. The same technique can be extended to model autocorrelations at longer time displacements, as described in Wilks (2006). From the general set of AR(n) models, the AR(2) model will shortly prove useful. Both AR(1) and AR(2) approaches are described in more detail in appendix B. The main aim of this section is to consider how best to estimate the inflation factor in practice, and to validate the normal statistical approaches against the experimentally-derived null distribution.

Before coming back to the populations of real forecast scores, these issues can be investigated using Monte-Carlo simulations. The analysis performed in Fig. 9 can be repeated with distributions made from normally-distributed random numbers (Fig. 14). As seen earlier (Tabs. 2 and 3 and Fig. 7) the paired differences of forecast scores have appreciable kurtosis, more outliers than might be expected, and high chi-squared values. It was found that a non-Gaussian PDF with such features could be simulated by shuffling together two distributions: a) putting 1,000,000 random numbers into a t-distribution with 8 degrees of freedom (using the t-distribution here simply as a convenient way of generating non-Gaussianity); b) to boost the tails of the PDF, 40,000 additional random numbers from a constant (rectangular) distribution between -4 and +4. This generated a population with a kurtosis of 1.3 and chi-squared of 36,399, similar to the real scores, though with the non-Gaussianity slightly exaggerated. However, the PDF and CDF of the means computed from 60-sample blocks was essentially no different from that predicted by the t-distribution (used conventionally, with 59 degrees of freedom). This is consistent with the central limit theorem: averages computed from a population are expected to be normally distributed even if the population distribution is not Gaussian itself. Hence the minor non-Gaussianity of the real forecast score differences can already be ruled out as the likely cause of the inflated error bars (and later results will confirm this.)

Figure 14 also examines temporal autocorrelation. The autocorrelation patterns representative of forecasts spaced 12 h and 24 h apart can be simulated by convolving a normally-distributed random number sequence with an appropriately chosen kernel. In this example, the autocorrelation between forecasts at day 5 has been taken from Fig. 4, and appendix B gives more details of how the simulation is done. This gives autocorrelation of 0.15 between neighbouring samples and 0.07 between every second sample, representative of a 12 h forecast spacing. This broadens the sampling distribution so that the confidence range would require need to be inflated to ± 2.5 in z-space, just as seen with real forecasts. The autocorrelation problem can be reduced by using forecasts separated by 24 h, and simulations have been made for this too, showing that an inflated confidence range of perhaps ± 2.1 would be required, in good agreement with the results of Fisher (1998), who also applied an AR(1) model to forecasts separated by 24 h. Hence the residual autocorrelation in the paired differences of forecast scores can explain the broadening of the null population in real forecast scores, e.g. in Fig. 9.

Table 4 shows the best available estimates of the inflation factor k using the autoregressive approach; see appendix B for more details. The lag-1 and lag-2 autocorrelation of the day-5 NH paired AMSU-A minus control differences from Fig. 4 has been used, based on the complete population of 1,866 to 1,885 samples (autocorrelations from the Noise-control population are not significantly different from the AMSU-A-control population). Using the AR(1) model applied to forecast scores with 24 h spacing requires an inflation factor of $k = 1.07$, very similar to the results of Fisher (1998). Since then, many scientists have moved to using forecasts every 12 h; applying the AR(1) model in this situation gives $k = 1.13$, which is an underestimate because it does not take into account the lag-2 autocorrelation. The AR(2) model gives $k = 1.22$, consistent with the results generated explicitly from the null population.

To further examine the agreement between the AR(2) results and the experimentally-derived null population, Tab. 5 shows estimates of z_{95} and k computed from this null population. The method illustrated

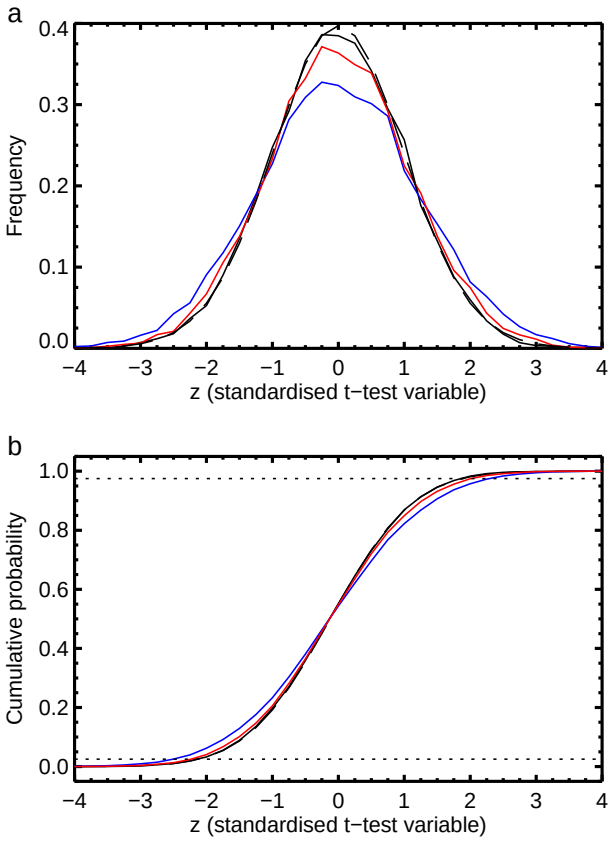


Figure 14: As Fig. 9, but based on randomly generated populations representing a true Gaussian (dashed), an uncorrelated non-Gaussian population with a kurtosis of 1.3 (solid black, barely distinguishable from the Gaussian), a Gaussian population with 0.07 autocorrelation, representing forecasts made 24 h apart (red) or with [0.15,0.07] autocorrelation, representing forecasts made 12 h apart (blue). Results are based on 16,666 non-overlapping blocks each containing 60 samples.

in Fig. 9 can be applied to different sample sizes and to the individual components (NH, SH, day-4 and day-10) as well as to the combined sample. However, to reduce quantisation in the results, the bin size used for computing the CDF is 0.01 here rather than 0.25 as shown in the figure. Further, assuming that with enough samples the distribution should be symmetric, z_{95} is estimated as the average of the the 0.025 and 0.975 quantiles. Nevertheless, the results are insensitive to these details (not shown). The main issue is that, as illustrated in Fig. 9, PDFs or CDFs computed from smaller numbers of blocks are strongly affected by sampling errors. These sampling errors have been estimated using a Monte Carlo technique, giving a 1-standard deviation range for the combined estimates of z_{95} and k. The range for the individual forecast days and hemispheres is not given explicitly but it can easily be estimated from those already given[2]. These sampling errors are large and they restrict what we can infer from these 'regional' results. This is why it is necessary to combine the four samples together. Except for those based on blocks of 10 samples, the estimates of k from the combined populations are all consistent with the $k = 1.22$ result from autoregressive methods.

Figure 4 shows that autocorrelations decline towards the end of the forecast range, whereas for the

[2]For example, to estimate the inflation factor k in the NH, day 10, there are 31 blocks available, similar to the 32 available when 230-day samples are combined across 4 populations, so the estimate and its confidence range would be roughly 1.09 ± 0.17.

Table 4: Estimates of k using autoregressive models and the autocorrelations for day-5 NH geopotential paired differences from Fig. 4, either for 12 h or 24 h separation between forecasts. A Monte-Carlo technique has been used to estimate the standard deviation of error in the k estimate - see appendix B.

		Autocorrelation		
Model	Forecast separation	Lag-1	Lag-2	k
AR(1)	24	0.07	0.0	1.07±0.02
AR(1)	12	0.15	0.0	1.13±0.02
AR(2)	12	1.15	0.07	1.22±0.04

Table 5: Estimates of z_{95} and k from the explicit null population. Estimates for '4 combined' use NH, SH day 4 and day 10 together to provide four times the number of blocks. The error in the combined estimates is generated by a Monte-Carlo method: normally distributed random numbers are passed through the same process that has been applied to the real forecast scores (with the same bin size etc.) and the standard deviation of 1000 possible estimates of z or k is computed.

z_{95}							
		Day 4		Day 10			
Samples per block	Number of blocks	SH	NH	SH	NH	Number of blocks	4 combined
10	186	3.00	3.12	2.38	2.25	744	2.59±0.09
30	62	2.44	2.78	2.42	2.51	248	2.50±0.13
60	31	2.27	2.59	2.94	2.18	124	2.36±0.16
120	15	1.98	2.48	2.35	2.12	60	2.46±0.23
230	8	1.59	1.26	1.87	2.34	32	2.34±0.34
k							
		Day 4		Day 10			
Samples per block	Number of blocks	SH	NH	SH	NH	Number of blocks	4 combined
10	186	1.32	1.40	1.08	0.96	744	1.14±0.04
30	62	1.19	1.36	1.18	1.23	248	1.22±0.06
60	31	1.13	1.29	1.47	1.09	124	1.18±0.08
120	15	1.00	1.25	1.18	1.07	60	1.24±0.12
230	8	0.81	0.64	0.95	1.19	32	1.19±0.17

AR(2) estimate of k, it was the day-5 autocorrelations that were used in the estimate. A reduced need for inflation at longer forecast ranges may be supported from the null population, at least when 10-sample blocks are used (Tab. 5): $k = 1.40 \pm 0.07$ at day 4 and $k = 0.96 \pm 0.07$ at day 10. However, with the larger sample sizes, for example with 60 samples per block, the difference is not significant: $k = 1.29 \pm 0.17$ at day 4 and $k = 1.09 \pm 0.17$ at day 10. Further, the estimate of k from small block sizes may be biased low, as suggested in appendix B. It could be speculated that the shorter block sizes do not fully represent all the temporal autocorrelation in the sample - perhaps if there is a very low level of, for example, intra-seasonal variation. These are interesting questions of detail, but only further study could resolve them. Broadly, the AR(2) estimate of the inflation factor based on day-5 autocorrelations is consistent with the estimate made non-parametrically from the null population.

There remains the issue of how the inflation factor should be determined practically. Appendix B shows that estimates of the lag-1 and lag-2 autocorrelations from typical population sizes are affected by sampling error and that this propagates into the k estimate. For example, an estimate of k based on autocorrelation computed from 360 samples would have a 1-standard deviation range of ± 0.09; for the 12 h data that could mean an error bar inflation that could easily fluctuate between 12% and 31% from one experiment to the next (as implemented in Verify, there is a hard minimum of 0% to prevent the error bars actually being deflated.) Hence, unless the sample size is very big, as in this study, it is not reliable to generate autocorrelation estimates from the population under test. Rather, a fixed estimate of k should be used. The best guess from this study for 12 h forecast separation would be $k = 1.22$, based on the explicit null population, which is supported both by the autocorrelation method and the block method. (For 24 h spacing, the recommendation is $k = 1.07$.) Looking at the tools available at ECMWF, Iver needs to start applying this inflation factor for the first time, and Verify, which already uses an AR(1) model, could either introduce an AR(2) method to cope with 12 h forecast scores, or alternatively the fixed inflation factors. If nothing else scientists should be aware that, in the presence of autocorrelation, a confidence range is only an estimate: error bars can themselves have error bars.

8 Multiple comparisons

It has been shown that when the small temporal autocorrelation in paired differences is taken into account, an individual t-test is statistically reliable. However, scientists often run multiple experiments and they usually look at more than one forecast score. So the issue of multiple comparison (multiplicity) will contribute to the excessive number of false or unrepeatable results in forecast testing. The 95% confidence range applies only to a single statistical test where it means there is only a 5% chance of a type-I error (i.e. incorrectly rejecting the null hypothesis). However, once many tests have been made, the chance of a false result becomes much greater. The issue of multiplicity is much better known in the medical sciences (e.g. Benjamini and Hochberg, 1995). In the atmospheric sciences it has only received attention where it is most obvious, when trying to compute a significance test for each location in a latitude-longitude field (e.g. Livezey and Chen, 1983; Wilks, 2006). In forecast verification, the problem is most apparent when trying to compute changes in forecast scores as a function of latitude and pressure. Geer et al. (2010) found it was necessary to make a Šidák-Bonferonni correction (e.g. Abdi, 2007) of the significance level from 95% to 99.8% to counter the tendency to assign excessive statistical significance across one latitude-height diagram. There, an assumption was made that there were about 20 independent statistical tests being made across the diagram. It has to be recognised that the number of independent statistical tests is generally much less than the number of tests being made, as neighbouring points in a spatial field are usually correlated. Much of the difficulty of dealing with multiplicity is in understanding this correlation between tests and correcting appropriately. Section 6 has shown that regional scores are

Table 6: The probability of a particular number of false results (Type-I errors) being generated by N significance tests computed using a 95% confidence interval. Generated from the binomial distribution following Wilks (2006). Bold type indicates the most probable number.

	0	1	2	3	4	5	6	7	8
N=4	**0.81**	0.17	0.01	0.00	0.00	0.00	0.00	0.00	0.00
N=8	**0.66**	0.28	0.05	0.01	0.00	0.00	0.00	0.00	0.00
N=12	**0.54**	0.34	0.10	0.02	0.00	0.00	0.00	0.00	0.00
N=16	**0.44**	0.37	0.15	0.04	0.01	0.00	0.00	0.00	0.00
N=32	0.19	**0.33**	0.27	0.14	0.05	0.02	0.00	0.00	0.00
N=112	0.00	0.02	0.06	0.11	0.15	**0.17**	0.16	0.13	0.09
N=124	0.00	0.01	0.04	0.08	0.12	0.16	**0.16**	0.15	0.11

correlated across the forecast range, so the same difficulties apply here.

The chance of a type-I error across multiple tests can be calculated using the binomial distribution as explained in Wilks (2006). Table 6 gives the probability of false results for certain numbers of tests. To apply this to forecast testing, imagine we want to evaluate two different scientific configurations against a control. From Sec. 6 there are about 4 different independent tests being made in the extratropical troposphere in each experiment: in the mid and long ranges, and in the SH and NH. As suggested by Sec. 6 we are only checking the RMS errors in 500 hPa height, and ignoring other vertical levels and other correlated scores like AC or wind error. This is a total of N=8 tests, and the table shows that at the 95% confidence level there is a 34% chance of one or more Type-I errors across this testing. Even with the relatively small number of independent tests being made when we compute hemispheric forecast scores, the chance of a false result can be much larger than expected.

We can return to the results from the samples containing 230 forecasts shown in Fig. 11b. With a 95% confidence range based on the correctly-inflated t-test, 1 of the 8 experiments showed 'statistically significant' results in the NH. This was perhaps unexpected given the naive view (valid only for a single statistical test) that we had a 5% chance of at least one Type-I error. In fact, there were 8 experiments and across the forecast range in the NH, 2 independent tests (perhaps day 3–6 and day 7–10). For those 16 independent tests, Tab. 6 suggests a 56% chance of at least one type-I error, or a 34% change of just one, so the results in Fig. 11b are understandable in this context.

To better assess the quality of agreement between results from the experimental null population and the statistical predictions, more samples are needed. The tests can be repeated for smaller samples, e.g. of 60 forecasts, where there are 31 separate blocks available. To ensure we are looking only at independent tests, only day 4 and day 10, SH and NH will be included, making 124 independent tests. Using the correctly-inflated confidence range there are 4 false results (not shown). At N=124 these 4 false results are within expected limits though towards the lower end of the expected range.

The suspicion that forecast testing generates an excessive number of false results is supported quantitatively by this study. It can be explained by two things: first, temporal autocorrelation must be correctly taken into account in the t-tests; second, the effect of making multiple comparisons needs to be recognised. As practical advice, the first and easiest step is to inflate the confidence limits of the individual tests as shown earlier. The multiple comparison issue could be addressed by adjusting the confidence limits further, taking into account the number of independent tests being made (the Šidák or Bonferonni approach) but this is technically and scientifically difficult. An approximate but still powerful technique is to use Table 6 as shown in this section. To recap, the first thing is to restrict the hypothesis testing to synoptic medium range scores (obviously the short-range, tropics and stratosphere are important too, but

due to systematic error and other effects, interpreting those scores is still quite a dangerous business, and needs to be done separately). N is obtained by counting up the number of experiments and multiplying by 4, allowing the usual four independent tests of NH, SH in mid and long-range. The number of significant changes in the experiments can be counted and compared to Table 6. Because runs of scores are correlated against forecast day, significant changes count only once in the mid-range (day 3-6) and once in the long range (day 7-10).

As an even simpler way to address the problem, it may be possible just to apply the following rules-of-thumb to medium-range scores:

- When a single NWP system experiment is compared to a control there is a 1 in 6 chance of at least one false result.
- For two experiments compared to a control, the chance is 1 in 4.
- Across three experiments compared to a control, the chance of a false result is 1 in 2.

This clearly highlights the danger of 'trawling' - of experimenting with a variety of different strategies for improving an NWP system and picking the strategy that gives the apparently significant result. If there is just one significant result (e.g. improvement in one experiment in the SH at days 3-6) that has a good chance of being false. However, if the multiple comparison problem is acknowledged and treated properly, results should be statistically reliable.

9 Minimum sample size needed for significance

It has been shown that paired differences in forecast quality between an experiment and control are dominated by forecast noise and that the real improvement from a typical scientific upgrade is a small signal on top of this. The correctly inflated confidence limits ($k = 1.22$) will be used to explore how this affects the size of sample required to verify a typical upgrade to an NWP system. The explicitly generated null population is valid for any experiment performed with a similar forecasting system, so based on the standard deviations of the null population (Tab. 2) it is possible to predict the size of the 95% confidence interval for any typical experiment in a high-quality NWP system.

Figure 15 shows the results. It is hard to achieve statistical significance in short experiments unless a very large change is made to the quality of the forecasting system. For example, with 40 forecasts, the confidence range at day 5 would be $\pm 2\%$. It is rare to see genuine improvements of this size in medium-range forecast skill. One of the most significant developments in ECMWF history, the introduction of 3D-Var, generated improvements in day 5 errors of 5% in the SH but just 1% in the NH (Andersson et al., 1998). The example of removing (or adding) a single AMSU-A instrument is more typical, with its impact of around 0.5%, and on average ECMWF scores are improving each year by about 2%, based on dozens of different scientific contributions. As mentioned before, the problem is that we are typically making small developments in what is already a very good forecasting system. Though it can be illustrative to make experiments where (for example) the whole satellite observing system is discarded, such experiments do not precisely address the question of how to make progress.

The 'typical' 0.5% impact will begin to show statistical significance at day 5 with a sample of 424 forecasts. This result suggests that even the recent ECMWF practice of demanding 6 months of testing for any proposed upgrade is likely to be insufficient (with two forecasts per day, this is about 360 forecasts in total). However, there is a greater problem of verifying scientific changes that have yet smaller impacts,

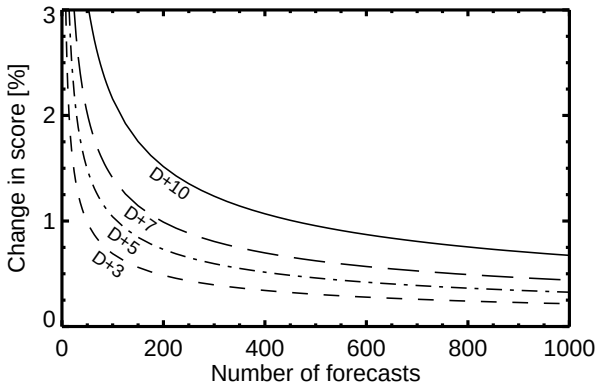


Figure 15: 95% confidence interval for 500 hPa geopotential height in the NH, expressed as a percentage of the RMS error of the control, and as a function of the number of forecasts in the sample used to compute the scores. If d_{95} is the interval read from this figure, the full confidence range is $\pm d_{95}$.

Table 7: In order to give statistically significant results at 95% confidence for a certain change in the NH 500hPa geopotential height forecast sores at day 5, the top section shows the number of samples required, and the bottom section the number of days of experimentation required. Results are shown using confidence limits inflated appropriately for forecast separations of 12 h, 24 h and 48 h.

Change in scores [%]	2	1	0.5	0.25
	Number of samples required			
Forecasts every 12 h ($k = 1.22$)	29	108	424	1690
Forecasts every 24 h ($k = 1.07$)	23	84	327	1301
Forecasts every 48 h ($k = 1$)	20	73	286	1136
	Number of days experimentation			
Forecasts every 12 h ($k = 1.22$)	14	54	212	845
Forecasts every 24 h ($k = 1.07$)	20	84	327	1301
Forecasts every 48 h ($k = 1$)	40	146	572	2272

particularly in preventing changes that make small degradations from entering the system. Even with 730 forecasts, a whole year of experimentation, at day 5 it would be impossible to distinguish a 0.2% improvement from a 0.2% degradation and in fact around 3000 forecasts would be required to give statistical significance. That is not a practical sample size: to create the 1,880 forecasts used for the current work required a major effort that lasted nearly a year, that took advantage of the installation of a new supercomputer, and that was suspended during busy periods. For the longer-range scores, such as day 10, a real 0.5% change in scores will not generate statistically significant results even with greater than 1,000 forecasts, as illustrated in Fig. 5. There seems no justification for extending the forecast range used in the verification of typical experiments beyond 10 days, since very few results will be statistically meaningful.

Given that the temporal autocorrelation between paired differences requires us to inflate the confidence interval, and this is worst when forecasts are made only 12 h apart, it might be thought better to verify forecasts spaced further apart, to minimise the autocorrelation and hence reduce the inflation required. However, this would be counterproductive, as shown in Tab. 7. Forecasts spaced 24 h apart still require a confidence interval inflated using $k = 1.07$. A 48 h spacing would probably remove all autocorrelation, which would allow the t-test to be used without inflation. The number of samples required to identify a

0.5% change with statistical significance would reduce from 424 to 327 or 286 respectively. However, this can only be obtained with 327 or 572 days of data assimilation, as opposed to 212 if the forecasts are made every 12 h. The best choice depends on the relative computational cost of the forecast versus the data assimilation, but usually the cost of data assimilation is substantially greater. Hence it is most efficient to generate forecasts every 12 h but to inflate the confidence range accordingly.

10 Conclusion

This study is a response to the difficulty of testing small improvements in forecast quality in the context of an already high-quality operational forecasting system. There is a natural desire among scientists to somehow 'go beyond' statistical significance and to draw inferences based on non-significant results. One aim of this study was simply to demonstrate how dangerous that is. The extent that chaotic variability affects forecast scores may not be fully appreciated. It has been shown how easy it is to generate apparently 'statistically significant' results by making a technical perturbation to a forecasting system, while keeping the science identical. However, even with statistical significance testing based on the Student's t-test, these happen more regularly than would be expected. In the differences between the technical perturbation ('Noise') experiment and the control we would hope to generate very few statistically significant results, in line with the aim to have only 5% chance of a false result (type-I error); in fact we would generate false results about 10 times too frequently, confirming the anecdotal evidence that we are getting too many unrepeatable results in forecast testing.

Significance testing suffers from two effects:

- If the residual autocorrelation in paired forecast differences is not taken into account, the Student's t-test is too generous. Two methods (first, an explicitly-generated null population, based on thousands of real forecasts and second, an AR(2) autoregression model) show that the 95% confidence range in z-space should be inflated by 22% when forecasts are separated by 12 h ($k = 1.22$ in Eq. 6; and for forecasts separated by 24 h, $k = 1.07$)
- When forecast scores are computed for different hemispheres and different time-ranges, as well as for different experiments, multiple hypothesis tests are being performed. The confidence range appropriate to a single test is not applicable to a whole family of tests, where every additional test adds to the probability of generating a false result.

The first effect is relatively easy to address and has the greater influence on reducing the number of false results. At ECMWF, the problem has become more serious since we started to verify forecasts every 12 h rather than every 24 h. Iver has not made any correction for autocorrelation and is most affected. Verify uses the AR(1) correction devised by Fisher (1998), but this underestimates the inflation factor needed for 12 h forecast separations (on average it estimates 14% rather than 22%) and, because the autocorrelation of the experimental population is used to derive the inflation factor, it is affected by sampling error (appendix B). For example some experiments will have had no inflation applied and for some the inflation will have been excessive. Hence both tools would best be updated to use fixed inflation estimates based on large populations of forecast scores, such as those presented here.

The second effect, the problem of multiple comparisons, is more difficult to handle. It is typical to generate an exhaustive set of forecast scores on many different levels and parameters, for different regions and views (e.g. zonal cross-sections and maps) and for different parameters, and this is vital for catching problems and for in-depth understanding of a forecast upgrade. Necessary as this is, it somewhat violates

the scientific philosophy of making the hypothesis before performing the test, and it runs the risk of generating spurious results through multiplicity. However, for medium-range synoptic scores (ignoring the short-range, the tropics and the stratosphere) there are only about 4 independent statistical tests being made in each experiment (in the mid-range, e.g. days 3–6, and the long-range, e.g. days 7 – 10, in the SH and NH). All dynamical scores are strongly correlated (whether computed as anomaly correlation or RMS, whether applied to geopotential, temperature or to the wind fields) and they are correlated across much of the troposphere. Changes in forecast scores are correlated from one forecast day to the next (even in a "Noise" experiment) so even if a run of a few days appears statistically significant, this still counts as just a single result. Even before thinking about multiplicity, this suggests it is sufficient to pick, in advance, a set of scores that will represent these four independent tests, and on which the most rigorous statistical evaluation is to be based. However, there are still enough independent tests being made to cause problems of multiplicity. For example when three experiments are being compared to a control, there are 12 independent tests of the medium-range synoptic scores being made. With a 95% confidence interval for the individual tests, that brings a 46% chance of getting one or more false results across all the synoptic testing. Hence, scientists should be sceptical of isolated results (for example, if forecast scores are improved only at day-4 in the SH in one of a set of three experiments, that should definitely be ignored). Section 8 gives guidance on how to treat this issue correctly, and presents simple rules of thumb that should prevent the worst misinterpretations of results.

It is sometimes assumed that because forecast scores are correlated in time, it is safer to verify only forecasts that are well-separated, for example to use only one forecast every 24 h, or indeed every 10 days. However, this is a poor strategy because is reduces the number of samples available. With forecasts every 12 h the confidence range needs to be inflated to take account of autocorrelation, but this is outweighed by the reduction in sampling uncertainty brought by the extra samples. It is always easier to detect changes in the quality of the forecasts using forecasts every 12 h than with a less frequent spacing.

A second aim of this study has been to give guidance on the sample sizes required to reliably test a small scientific change. The most significant developments in ECMWF history have demonstrated impacts on medium-range forecast scores of around 1% to 5%, but something much smaller is typical. The example here has been the 0.5% change in scores resulting from the addition or removal of a single AMSU-A instrument (where 6 others are still assimilated). A steady flow of smaller changes like this has contributed significantly to the improvement of scores over the last few decades, and in any case, even the minor developments need scientific testing to prove they are not making scores worse. But to reliably test a 0.5% change requires around 400 forecasts, which is computationally very expensive. In fact, if the recommendations of sample sizes in this report were followed unthinkingly, at ECMWF the supercomputer would become so congested that no-one would get any work done.

In the situation where traditional medium-range scores cannot always be used as a guide to the quality of scientific developments in forecasting, what do we do? We could just assume a-priori that something that improves the science will improve the forecasts. However, bugs can creep into the system and science can be less well-understood than anticipated, so it is not realistic to abandon testing altogether. It is probably best to test fewer configurations but to do it more rigorously, generating many more samples and accepting the additional cost. A particularly bad strategy is that of making various minor changes to the system in various different experiments and then trying to decide between them using the forecast scores. This is basically 'trawling', and sets up a situation where the effect of multiplicity will generate misleading results. One of the candidate developments may be chosen almost at random. Instead, it is better to carefully assess those minor changes before starting long forecast runs, and then to take just the best candidate forward into a full experiment. For example, when choosing among three different improved cloud screening schemes, it should be possible to find detailed evidence to distinguish the best

among them (perhaps, in this case, by comparing to independent cloud retrievals.) Then the hypothesis that better cloud screening improves forecast scores can be taken forward in a single, long, high-quality forecast experiment.

There are a number of other strategies that may help. First, ensembles are one way to generate additional samples, though there are statistical challenges in interpreting the results and in generating confidence intervals. Second, in some cases, it might be appropriate to increase the gap in model resolution used in testing versus that of a full operational system. However, the applicability of results from relatively low resolution compared to those from a full-resolution model requires confirmation. A third possibility is to concentrate on short-range forecast testing, where statistical significance is apparently easier to achieve. However, there are many pitfalls in short-range verification, particularly when using analyses as the reference, so often observation-based verification can be more reliable. Another issue with the short-range strategy (from the point of view of medium-range forecasting) is that it assumes that improvement in forecasts in the very short range will actually lead to improvement in the medium range. That is clearly true in general, averaged over many cases, but if it is only very small areas (whose locations change continuously) that control the growth of baroclinic errors, these areas are likely to get lost in hemispheric scores. Finally, innovative methods for diagnosing observation impact such as adjoint sensitivity (e.g. Baker and Daley, 2000) or ensemble spread (e.g. Harnisch et al., 2013) could be useful but because they are usually only applied to short-range verification, they face exactly the same issues. Short-range forecast verification has been excluded from the current study, but it is of great interest to understand it better. Overall, medium-range dynamical forecast scores remain the simplest, most direct and reliable tools for evaluating scientific improvements, but they do require sufficiently large samples along with rigorous statistical evaluation.

Acknowledgments

The following are thanked for assistance, discussions and reviews: Mike Fisher, Michael Rennie, Martin Janoušek, Elias Hólm, Stephen English and Erland Källén. Also, Tomas Wilhelmsson and Deborah Salmond are acknowledged in their thankless task of controlling the numerical reproducibility of the forecasting system, which was essential in creating the noise experiment.

A Results using operational analysis as the reference

The results presented have been based on own-analysis verification. However, the entire study has been repeated using ECMWF operational analyses as the verification reference. The results are generally very similar and lead to very similar conclusions. Results and figures in Secs. 3, 4 and 6 are nearly identical except for the changes in forecast scores in the early range (day 2 and earlier) of Fig. 5. Later sections rely on results that are themselves affected by sampling noise in their finer details. So where results are based on a limited number of blocks, details of the figures do differ with operational analysis as the reference, but all the conclusions are the same, including that the t-test is overly generous without an inflation of the confidence range of about $k = 1.22$.

B Modelling variance inflation with autoregression models

Fisher (1998) and Wilks (2006) show how to specify an inflation factor $k = \sqrt{V}$ in Eq. 6 to account for autocorrelation, under the condition that the autocorrelation can be modelled using an autoregression (AR) model. The day-5 NH 500 hPa geopotential height forecast score differences have a 12 h autocorrelation $r_1 = 0.15$ and 24 h autocorrelation $r_2 = 0.07$ (Fig. 4). Modelling this autocorrelation with an AR(2), V can be computed using:

$$\begin{aligned} \phi_1 &= (r_1(1-r_2))/(1-r_1^2) && (7)\\ \phi_2 &= (r_2-r_1^2)/(1-r_1^2) && (8)\\ \rho_1 &= \phi_1/(1-\phi_2) && (9)\\ \rho_2 &= \phi_2+\phi_1^2/(1-\phi_2) && (10)\\ V &= (1-\rho_1\phi_1-\rho_2\phi_2)/(1-\phi_1-\phi_2)^2 && (11) \end{aligned}$$

The AR(1) solution can be obtained by setting $r_2 = 0$. A more general solution for any AR(N) can be found in Wilks. For the NH example, $k = \sqrt{V} = 1.22$ which is quite similar to the implied inflation from the revised confidence limits suggested in Sec. 5. Rather than compute a null population explicitly, as is done in this work, it is possible to estimate the inflation factor directly from autocorrelation estimates of the population of paired forecast score differences, e.g. Fig. 4. This would potentially enable the inflation factor to adapt flexibly to different forecast-spacings (e.g. every 12 h versus every 24 h) and to take into account the lessening of autocorrelation at longer forecast ranges that is seen in Fig. 4. The main problem is that estimates of autocorrelation are themselves subject to sampling uncertainty. Table 8 explores the effect of this on estimates of k using a similar Monte-Carlo method to that used to estimate the error bars for Fig. 4. For the typical sample sizes used in forecast verification, this will add considerable uncertainty to the estimates of the error bars themselves. Only for very large samples, such as with the 1,885 forecasts available in this study, does the estimate of k become reliable enough. Another interesting effect is that when small sample sizes are used, even beyond the large sampling error, the estimate of k has a clear low bias; this parallels the low bias in the non-parametric estimates of k in Sec. 7. A possible explanation is that the samples at the beginning and end of the block may contribute less to the autocorrelation and hence start to dominate the results.

C Using bootstrapping to sample the null distribution

The results in the main study are limited by the number of samples that can be generated for comparison to the t-distribution whilst requiring contiguous and non-overlapping groups of forecasts. A larger set of samples could be computed using the moving block bootstrap resampling technique described by Wilks (2006). The basic unit is a small contiguous block of paired differences which is long enough to preserve most of the autocorrelation in the data but short enough to allow the population to be resampled in many different combinations. Here, a block length of 4 forecasts is used, which can be justified following Wilks (1997), based on the 0.15 autocorrelation between subsequent paired differences (from Sec. 3). These blocks are randomly drawn from the null population, with replacement, and assembled to create a sample of data from which a mean can be computed. Also, to gain the largest and cleanest sample, the four independent populations (NH and SH at day 4 and day 10, as used in Sec. 4) can be used. From each of the four null populations, 10,000 samples of size n are created by bootstrapping and their mean $\overline{d}$ and standard deviation s are computed. These can be transformed into z-space via Eq. 5. The results are shown in Fig. 16.

Table 8: Reliability of inflation factor k generated from autocorrelation parameters r_1 and r_2, which themselves have been estimated from a sample of finite size. A Monte-Carlo method is used: sequences of normally distributed random numbers are made autocorrelated by applying a convolution function; r_1 and r_2 are estimated and from these k is computed. 1000 randomly generated blocks of 10, 60, 360 or 1885 samples have been used.

Forecasts every 12 h (simulated by [0.025, 0.065, 0.82, 0.065, 0.025] convolution)				
Sample size	Mean	Std. dev.	Min	Max
10	0.869	0.324	0.22	2.14
60	1.161	0.219	0.65	1.98
360	1.214	0.090	0.96	1.59
1885	1.223	0.038	1.11	1.35
Forecasts every 24 h (simulated by [0.035, 0.93, 0.035] convolution)				
Sample size	Mean	Std. dev.	Min	Max
10	0.811	0.305	0.21	2.02
60	1.025	0.190	0.56	1.77
360	1.065	0.077	0.85	1.37
1885	1.071	0.033	0.98	1.17

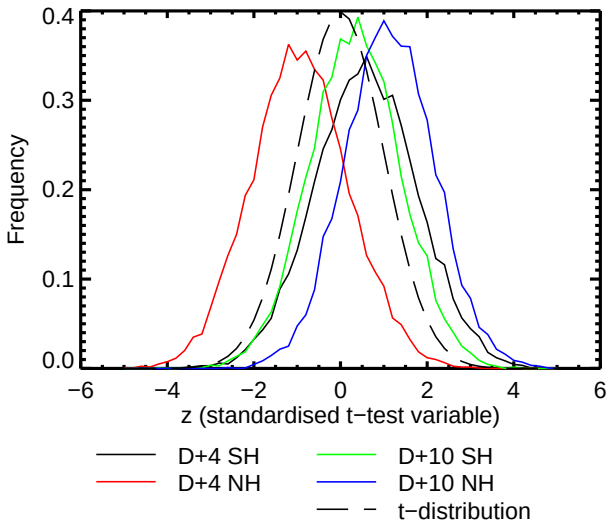


Figure 16: (a) PDF of z-tests generated from bootstrapped samples of 720 forecasts (solid) and the Student's t-distribution (dashed). Based on a total 10,000 for each of the four independent populations.

However, it looks like even bootstrapping cannot beat the limitations of a relatively small sample. For example, at day 10 in the NH, even over 1,885 forecasts, the Noise experiment has RMS errors that are on averge 0.35 m larger than in the Control (Tab. 2). This must be a purely random effect, as the experiments have no scientific differences. Bootstrapping creates a smoother histogram based on samples from these 1,885 forecasts, but it can do nothing about these biases. Distributions like those in Fig. 9 can be generated but, although they are smoother, for the larger-size samples they are generally just as lopsided. Removing the bias would be incorrect, as that is part of the random variability. There does not seem much to be gained from bootstrapping for the present work.

References

Abdi, H. (2007). The Bonferroni and Šidák corrections for multiple comparisons. In N. Salkind (Ed.), *Encylopedia of Measurement and Statistics*, pp. 103–107. Sage.

Andersson, E., J. Haseler, P. Undén, P. Courtier, G. Kelly, D. Vasiljevič, C. Brankovi, C. Cardinali, C. Gaffard, A. Hollingsworth, C. Jakob, P. Janssen, E. Klinker, A. Lanzinger, M. Miller, F. Rabier, A. Simmons, B. Strauss, J.-N. Thépaut, and P. Viterbo (1998). The ECMWF implementation of three-dimensional variational assimilation (3D-Var). III: Experimental results. *Quart. J. Roy. Meteorol. Soc. 124*, 1831–1860.

Baker, N. L. and R. Daley (2000). Observation and background adjoint sensitivity in the adaptive observation-targeting problem. *Quarterly Journal of the Royal Meteorological Society 126*(565), 1431–1454.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. Series B*, 289–300.

Bonavita, M., L. Isaksen, E. Hólm, and M. Fisher (2015). The evolution of the ECMWF hybrid data assimilation system. *Quart. J. Roy. Meteorol. Soc. doi:10.1002/qj.2652*.

Fisher, M. (1998). Statistical significance testing of forecast experiments. *ECMWF internal memorandum R48.8/MF/30 issued 23rd April 1998; available as a PDF from the author*.

Geer, A. J., F. Baordo, N. Bormann, and S. English (2014). All-sky assimilation of microwave humidity sounders. *ECMWF Tech. Memo., 741, available from http://www.ecmwf.int*.

Geer, A. J., P. Bauer, and P. Lopez (2010). Direct 4D-Var assimilation of all-sky radiances: Part II. Assessment. *Quart. J. Roy. Meteorol. Soc. 136*, 1886–1905.

Harnisch, F., S. Healy, P. Bauer, and S. English (2013). Scaling of GNSS radio occultation impact with observation number using an ensemble of data assimilations. *Monthly Weather Review 141*(12), 4395–4413.

Hodyss, D. and S. J. Majumdar (2007). The contamination of 'data impact' in global models by rapidly growing mesoscale instabilities. *Quart. J. Roy. Meteorol. Soc. 133*, 1865–1875.

Isaksen, L., M. Bonavita, R. Buizza, M. Fisher, J. Haseler, M. Leutbecher, and L. Raynaud (2010). Ensemble of data assimilations at ECMWF. *ECMWF tech. memo. 636*, available from www.ecmwf.int.

Jolliffe, I. T. and D. B. Stephenson (2012). *Forecast verification: A practitioner's guide in atmospheric science* (2nd ed.). Wiley-Blackwell.

Livezey, R. E. and W. Y. Chen (1983). Statistical field significance and its determination by Monte Carlo techniques. *Mon. Weath. Rev. 111*, 46–59.

Lorenz, E. (1982). Atmospheric predictability experiments with a large numerical model. *Tellus 34*(6), 505–513.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *J. Atmos. Sci. 20*(2), 130–141.

Magnusson, L. and E. Källén (2013). Factors influencing skill improvements in the ECMWF forecasting system. *Mon. Weath. Rev. 141*(9), 3142–3153.

Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Weath. Rev. 116*(12), 2417–2424.

Radnoti, G., P. Bauer, A. McNally, C. Cardinali, H. S., and P. de Rosnay (2010). ECMWF study on the impact of future developments of the space-based observing system on numerical weather prediction. *ECMWF Tech. Memo., 638, available from http://www.ecmwf.int*.

Robel, J. (2009). NOAA KLM user's guide (February 2009 revision). *Available from http://www.ncdc.noaa.gov/oa/pod-guide/ncdc/docs/intro.htm*.

Simmons, A. and A. Hollingsworth (2002). Some aspects of the improvement in skill of numerical weather prediction. *Quart. J. Roy. Meteorol. Soc. 128*(580), 647–677.

Thorpe, A., P. Bauer, L. Magnusson, and D. Richardson (2013). An evaluation of recent performance of ECMWF's forecasts. *ECMWF Newsletter* (137), 15–18.

Wilks, D. (1997). Resampling hypothesis tests for autocorrelated fields. *J. Clim. 10*(1), 65–82.

Wilks, D. (2006). *Statistical methods in the atmospheric sciences* (2 ed.). Academic Press.