How detectable are improvements in forecast quality?

Technical content: high
SPECS Fact sheet #5
April 2015

The big question. Numerical models of the atmosphere undergo constant improvements, for example by increasing the resolution, including new feedbacks, or simply fixing bugs in the source code. It is often not clear whether such an improvement of the model has resulted in an improvement in forecast quality. Alternatively, there might be two models provided by two different climate centres, and it is of interest which of the two provides better forecasts. The big question is often **"Does model A produce better forecasts than model B?"**. This question can be addressed by statistical analysis of past forecasts (hindcasts).

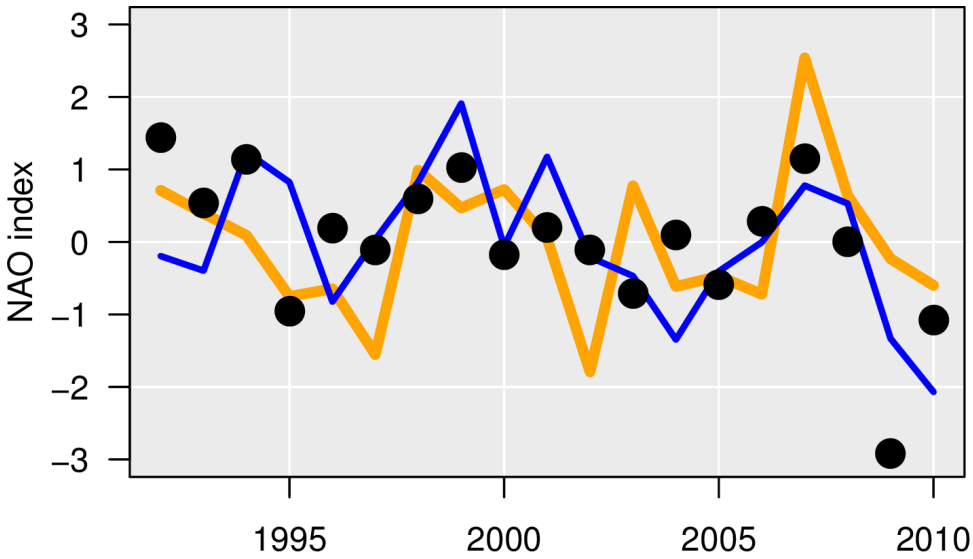

Figure 1: 19 years of ensemble forecast means of the "old" (orange) and the "new" (blue) forecast model, and the verifying observations (points).

The North Atlantic Oscillation (NAO) is known as one of the major drivers of European climate. The NAO index is also a cherished prediction target of climate scientists to test the capabilities of seasonal-to-decadal forecast models. Consider the two ensemble mean hindcasts for 19 years of winter NAO index shown in Figure 1: The "old" model is based on an ensemble of 15 forecasts, and the "new" model generated a 24-member ensemble. Both sets of forecasts are initialized in early November. Upon visual inspection it is not clear which of the forecasts should be preferred. The correlation coefficients with the observation was found to be 0.38 for the "old" model, and 0.56 for the "new" model, so there is an indication that the "new" model constitutes an improvement over the "old" model. **But might the difference of 0.18 simply be due to random sampling variations?**

This project is funded by the Seventh Framework Programme (FP7) of the European Commission GA 308378
Design by IC3
specs@ic3.cat

Testing for improvement in correlation. In order to test this, it is common practice to formulate the null-hypothesis (H0) that, taken over infinitely many hindcasts, the two systems have the same correlation. We then ask, if H0 were true, how frequently would the difference in sample correlation, taken over a random sample of 19 hindcasts, exceed the observed value of 0.18? This hypothetical frequency is called the p-value of the test. A small threshold, called the significance level α (common values for α are 0.01, 0.05, 0.1), is defined, below which a p-value is deemed "significant": If the p-value is smaller than α , the observed difference in correlation is considered too large to be compatible with H0, and H0 is therefore rejected.

In our example, the p-value of a test for zero difference in correlation is 0.23 (Steiger, 1980). That is, under the null-hypothesis, there is a 23% chance of observing a difference in sample correlation at least as large as the 0.18 that was observed for the two NAO hindcasts. Even though the new model has higher sample correlation than the old model, a correlation difference of 0.18 or more would be observed 23% of the time if H0 were true. **We cannot be sure whether the new model forecast is really an improvement over the old model forecast, or whether the observed difference in correlation is just due to random sampling variations.** The p-value of the observed difference in correlation is too large to be deemed "significant", and therefore H0 is not rejected. Formally, no improvement is detected by the test.

	H0: systems have the same skill	**H1: the new system is better**
do not reject H0	true negative	false negative
reject H0	false positive	true positive

Table 1: Null-hypothesis, alternative hypothesis and the two possible decisions

The significance level α controls the probability of a false positive, i.e., of falsely detecting a difference in correlation when there is none. If the correlation has not improved, and we conduct a test at significance level 0.05, we will be deceived into thinking that there has been an improvement about 5% of the time. But if we want to know how detectable an actual improvement in correlation is, we should analyse the probability of a true positive, i.e., of a correct detection, rather than the probability of a false detection.

UNIVERSITY OF EXETER ECMWF Met Office IC3

This project is funded by the Seventh Framework Programme (FP7) of the European Commission GA 308378

Design by IC3
specs@ic3.cat

How detectable are improvements in forecast quality?

Statistical power analysis. Let us assume that the correlation coefficients of the two forecasting systems, taken over an infinitely large hindcast dataset, were indeed 0.38 and 0.56. This would mean that the new forecast is indeed "better" than the old forecast, and a statistical test that does not reject the null-hypothesis of zero difference would commit false negative error. An important question is "What is the chance of a true positive?", i.e. how frequently does the test correctly reject the null-hypothesis of zero correlation? The probability of a correct detection is called the power. **What is the power of the test?**

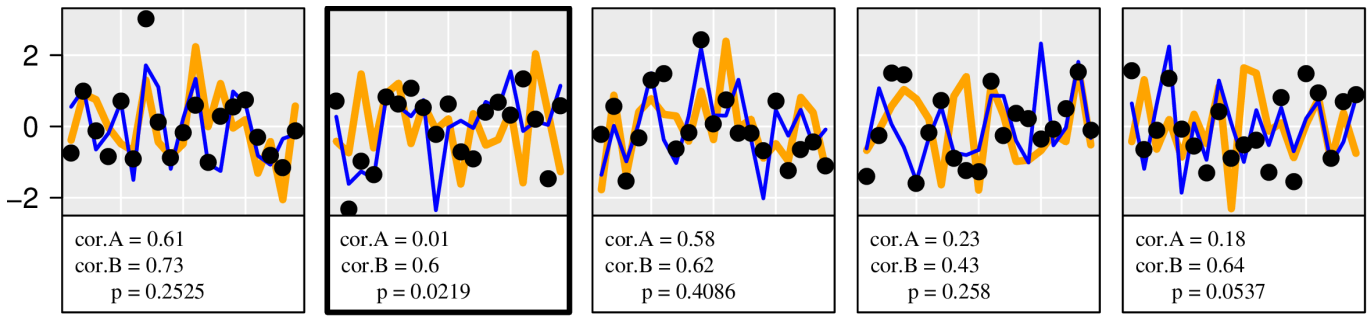

Figure 2: Artificial data similar to the original NAO hindcast data of Figure 1. In all 5 examples, forecast B has higher correlation than forecast A, but only in one out of five is the difference significant at the 5% level.

One approach to address this question is to fit a statistical (multivariate Normal) model to simulate artificial hindcast data. Taken over infinitely many samples, the artificial data has the correlation coefficients above, but taken over a small data set, correlations vary due to sampling variations (see Figure 2). The power of the test is the percentage of data sets in which the observed difference in correlation is found to be significant. For this data, using a significance level of α=0.05, the hypothesis of zero correlation difference is correctly rejected only about 18% of the time. **That is, there is a chance of over 80% that the improvement from the old to the new model remains undetected.**

The above result can be considered representative of most comparative forecast verification in seasonal-to-decadal climate prediction. The sample size and the differences between the competing forecasts are usually too small to produce significant test results. **In conclusion, any improvements of seasonal-to-decadal predictions are unlikely to be detected by statistical tests.**

This project is funded by the Seventh Framework Programme (FP7) of the European Commission GA 308378

How detectable are improvements in forecast quality?

Discussion:

How does the example relate to practical model development? Climate model development is an **incremental process**, where small changes are constantly applied to the forecast system. Small changes supposedly lead to only small improvements of forecast quality. The increase in correlation from 0.38 to 0.56 is larger than what would normally be expected, so the statistical power of most comparative hindcast experiments might be even lower. Furthermore, as forecast models get better and approach the inherent predictability limit of the climate system, **future improvements will naturally become smaller**, and thus harder to detect.

How can we increase our ability to detect improvements in models? Statistical power depends on sample size, internal variability of the data, the statistical testing method, and the desired significance level. While there is little control over internal variability of the forecast and observation data, and increasing sample size (number of years) is computationally expensive, **the verification measure, the statistical testing method, and the significance level might be chosen to maximise the power of detecting improvements** (e.g. Wilks, 2010). Higher power might be also obtainable by testing for improvements at many grid points or using multiple forecast targets, applying suitable adjustments for **multiple testing** (Ventura 2004, Wilks 2006). A **formal decision analysis** could take into account not only the uncertainty about the difference in forecast quality, but also possible costs of falsely rejecting a superior model, or falsely replacing a skillful model by a less skillful one.

What other aspects of model performance need to be considered? So far we have considered only model predictive skill. **Physical realism**, i.e., the ability of the model to mimic general features of the real world, independent of predictive skill, is also an important factor that guides the development and improvement of climate models. A new climate model that violates basic physical principles such as conservation of energy might not be preferable, even if its apparent forecast performance has improved. Similarly, even if no statistically significant improvement of hindcast skill can be detected, one model might be trusted more because its representation of key aspects of the climate system is more realistic. **Combining results from statistical inference and physical reasoning** is an important part of the model development process.

This project is funded by the Seventh Framework Programme (FP7) of the European Commission GA 308378

How detectable are improvements in forecast quality?

The multi-model approach

As we have seen, the low power of statistical tests leads to an **undecidability problem** - we can never be sure which model will provide the best forecasts for a particular climate mode in the future. So, how can we make the best of the international pool of models? This conundrum has motivated a **multi-model approach** in which a consensus forecast is generated by combining the ensembles from a number of candidate models. A multi-model approach eliminates the need to look for the best model, but it requires simultaneous investments in developing and running a collection of climate models. (see SPECS Fact sheet #4: Climate prediction with multiple sources of information)

References:

•Steiger (1980): Tests for comparing elements of a correlation matrix. doi: 10.1037/0033-2909.87.2.245

•Ventura, Paciorek and Risbey (2004): Controlling the Proportion of Falsely Rejected Hypotheses when Conducting Multiple Tests with Climatological Data. doi: 10.1175/3199.1

•Wilks (2006): On "Field Significance" and the False Discovery Rate. doi: 10.1175/JAM2404.1

•Wilks (2010): Sampling distributions of the Brier score and Brier skill score under serial dependence. doi: 10.1002/qj.709

This project is funded by the Seventh Framework Programme (FP7) of the European Commission GA 308378