All models are wrong, some can be made less wrong

Roman Binter[1] and Leonard A. Smith (Email: *lenny@maths.ox.ac.uk*)[1,2]

[1] Centre for the Analysis of the Time Series, London School of Economics; [2] Pembroke College, Oxford

Abstract

In the presence of noise and imperfect understanding of a forecasted system, models are destined to be wrong. Given a *core* forecasting model Φ, [2] suggest correcting for the systematic error in the core model by applying a second model Ψ which aims to minimize the root mean square error of the $\Psi\Phi$ forecast. In this poster, we revisit their approach and consider both the aim and the means of this suggestion. First we suggest a new approach that iteratively corrects the core model Φ. We call this approach the Predictor-Corrector (PC). Then we follow by a discussion of whether or not RMS error represents effective means of evaluation in nonlinear forecasting. We conclude that it does not.

1 Additive vs. Iterative Corrector

When attempting to correct for a systematic model error of a core model Φ, [2] suggest to construct a separate model Ψ, which estimates the in-sample systematic error at each lead time. In the forecasting mode, the Ψ model is used to provide a forecast of the out-of-sample error. The error forecast is then added to the forecast produced by the core model Φ, yielding a corrected forecast $\Psi\Phi$. The basic idea is illustrated in the left panel of Fig: 1. The core model Φ forecast (green) is corrected by in-sample error estimate (red arrows) to obtain the corrected $\Psi\Phi$ forecast (blue). The Ψ model can be constructed in a number of ways. In this poster we choose two different Ψ models used in [2]. The first is based on the linear model while the second makes use of Radial Basis Functions.

An alternative to $\Psi\Phi$ approach is to use a separate model to estimate the in-sample systematic error of 1-step ahead forecast only, and then keep correcting forecasts produced by the core model Φ at each step. We illustrate the idea in the right panel of Fig: 1. Consider, an initial observation being input into the core model Φ to generate leadtime 1 forecast (green line on the $(0, 8)$ interval). Using a forecast of a 1-step ahead systematic error (red arrow at $x = 8$), produced by a sepparate model, we correct the 1-step ahead forecast. As a result we obtain a leadtime 1 *PC* forecast (blue). The leadtime 1 PC forecast is then directly input into the core model to produce leadtime 2 forecast.

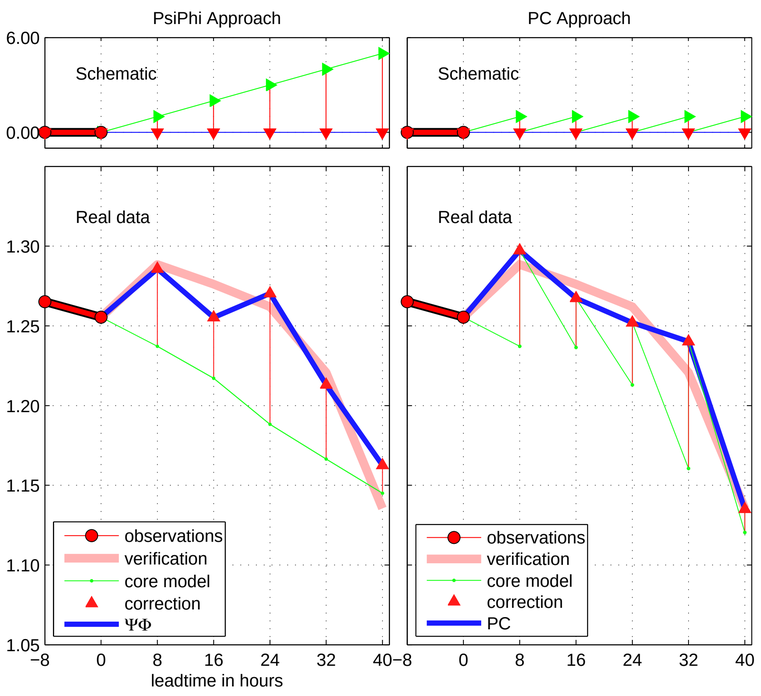

Figure 1: **Contrasting $\Psi\Phi$ and PC approaches:** Left top panel shows a schematic of the $\Psi\Phi$ approach. An iterative forecast of the core model Φ (green arrows) is initialized at a current observation (red dot) and produces forecasts for all leadtimes. In the next step the Ψ model produces forecasts of the systematic error (red arrows) for each leadtime based on the output of the Φ model. The error estimates are then added to the Φ forecast to obtain final $\Psi\Phi$ forecast (blue vertical line). Top right panel shows the PC approach, where one-step ahead forecast (1st green arrow) of the core model is immediately corrected to obtain final PC forecast at the leadtime 1. The leadtime 1 forecast is then input into the core model producing leadtime 2 forecast (2nd green arrow), which is immediately corrected yielding PC forecast at leadtime 2. This process is repeated for required number of leadtimes. The bottom panels demonstrate the two approaches on real data, where they forecast 40 hours of the Lorenz 84 system. The pale red line represents the verifications.

2 Forecast evaluation

The value of a forecast can be quantified by comparing it with a reference forecast. In this poster, skill is reported using the *log p* score called ignorance *Ignorance* (IGN) [5], which is the only proper, local skill score [1] for continuous distributions. Ignorance is defined as $I = \frac{1}{T}\sum_{t=1}^{T} -log_2(p(y_t))$, where $p(y_t)$ is a probability assigned by the model to the verification. The mean is taken over an archive of past forecast-verification pairs recorded at time $t = 1, \ldots, T$. *Relative ignorance* determines skill relative to some reference forecast, $I_R = I_{forecast} - I_{reference}$, by convention the lower the value of IGN the better the forecast. Using *log* with base 2 expresses ignorance in bits: if the relative ignorance is -1 then on average the forecasting model places twice (2^1) as much probability on the verification as the reference. Likewise, $I_R = -2$ means that the reference forecast puts 2^2, or 4 times more, while $I_R = 0.5$ would mean $sqrt(2) \sim 1.4$, or about 40% more.

3 Forecasting Lorenz 84/63

We have applied both the PC and the $\Psi\Phi$ approaches to several systems. In this poster we present forecasting results for the x-variable of Lorenz 84 system [4] and the Lorenz 63 system [3]. The Lorenz 84 system is sampled at 8 hours of the system time with the noise level set to 0.5% of the range of a given variable. Lorenz 63 system is sampled at 0.032 seconds and is also obscured with the noise level of 0.5% of the range.

In both cases we have used training, evaluation and testing sets of equal size. The testing sets consist of 1200 segments of significant length: 100 days (300 leadtimes) in the case of Lorenz 84 and 10 seconds (also 300 leadtimes) in the Lorenz 63 case.

We first focus on the Lorenz 84 system and use 5 different models to forecast the evolution of the x variable. The models are: Perfect model (PMS), core model (Φ), predictor-corrector (PC), $\Psi\Phi$ model based on linear model ($\Psi\Phi$ LSQ) and $\Psi\Phi$ model based on radial basis functions ($\Psi\Phi$ RBF). Note, that the PMS is used namely for benchmarking purposes.

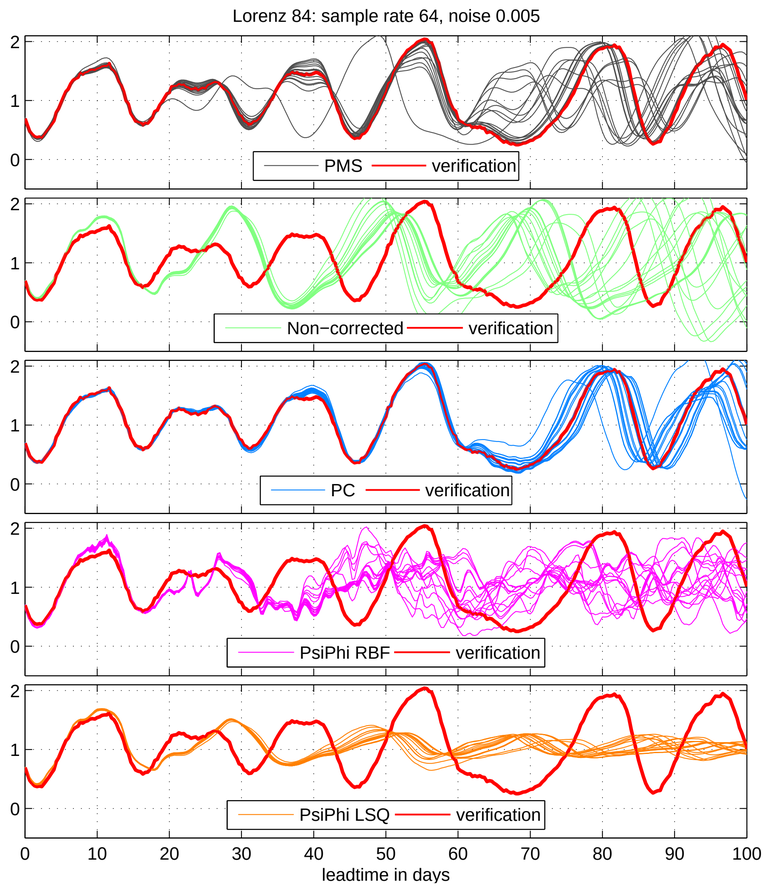

Figure 2: **Lorenz 84 forecasts:** From top to bottom, PMS, non-corrected core model Φ, PC, $\Psi\Phi$ RBF and $\Psi\Phi$ LSQ forecasts of a selected 100 hour long segment of the Lorenz 84 system. The PC outperforms all the other 4 (imperfect) models. The $\Psi\Phi$ LSQ quickly resorts to forecasting the mean of x, the non-corrected Φ model looses skill after 20 hours. The $\Psi\Phi$ RBF marginally improves the particular forecast.

In Fig: 2 we present results for a selected forecast for illustration. In this particular case we see that the non-corrected core model looses forecasting skill after about 15 hours. The PC method improves the forecast massively as several ensemble members stay close to the verification (red line) for all 100 hours. The $\Psi\Phi$ RBF also positively improves the core model's forecast but not as significantly as PC. The $\Psi\Phi$ LSQ correction have negligible impact, the forecasts quickly converge to the mean.

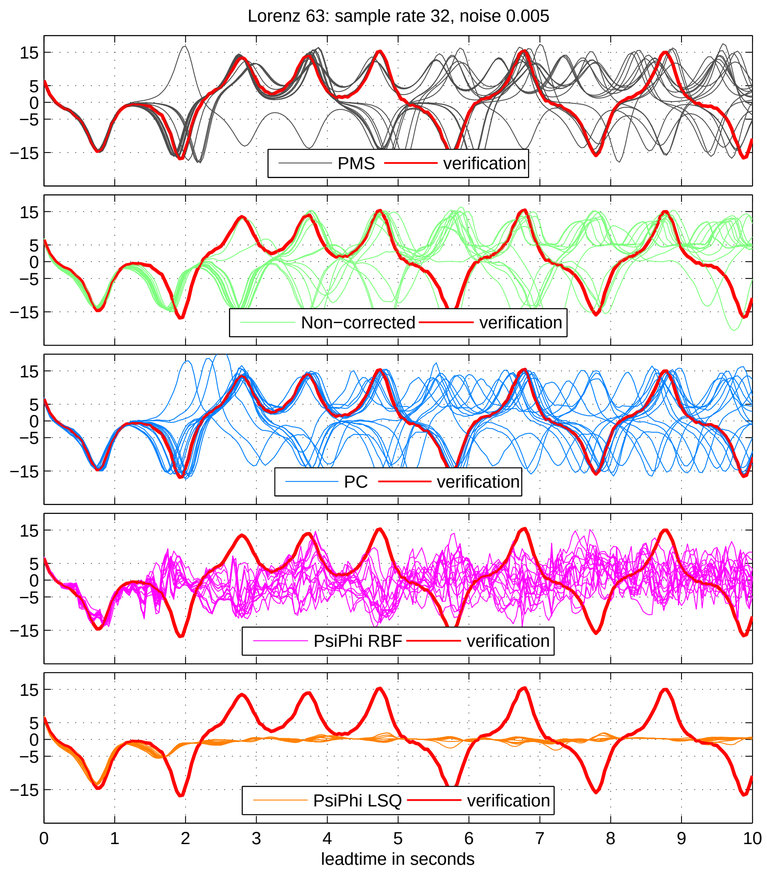

Figure 3: **Lorenz 63 Forecasts:** Same models as in Fig: used to generate forecasts of x variable of the Lorenz 63 system 10 seconds ahead. The non-corrected model looses skill after 1.5 seconds. PC performs very well, while $\Psi\Phi$ LSQ converges to mean rapidly.

Next, we look at the forecasts of the Lorenz 63 system, Fig: 3. The $\Psi\Phi$ LSQ forecast converges to the mean after 1.5 seconds. The $\Psi\Phi$ RBF somewhat converges to the mean although the effect is not pronounce. The PC performs very well.

4 Evaluating the correctors:

In Fig: 4 we evaluate the Lorenz 84 forecasts using *Relative Ignorance*. The non-corrected Φ model (green) looses skill before leadtime of 5 days. The Ignorance of the $\Psi\Phi$ LSQ shows that the linear based corrector does not really improve the core model. The PC produces very good corrections maintaining skill up to 60 days, i.e. about double of the $\Psi\Phi$ RBF and 10 times longer then the non-corrected Φ model.

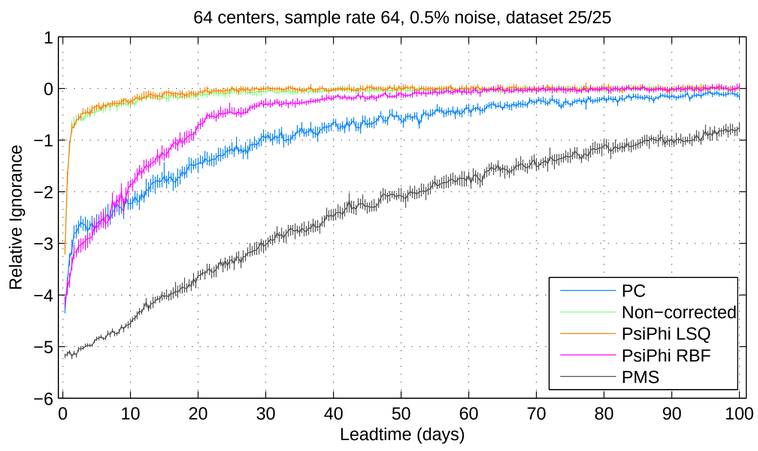

Figure 4: **Evaluation by Ignorance.** Based on Ignorance the best performer of the 4 imperfect model is the PC (blue) followed by the $\Psi\Phi$ RBF (magenta) which has also significantly improved the core model and at very short leadtimes up to 2 days outperforms the PC. Both the non-corrected Φ model and the $\Psi\Phi$ LSQ quickly loose forecasting skill beyond leadtime of 2-3 days. The Ignorance of PMS (gray) is shown for comparison.

5 RMS evaluation can be misleading

Scoring rules are used to (a) evaluate the forecasting model performance and to (b) select the best performer if several models are available. Despite the analytically proved drawbacks of the root mean square error (RMS) in non-linear setting [1], RMS remains to be frequently used when evalulating non-linear forecasts. In the following we demonstrate that when used for model selection in non-linear applications, the RMS may misslead the forecaster into choosing a model that is more wrong than it's alternative(s).

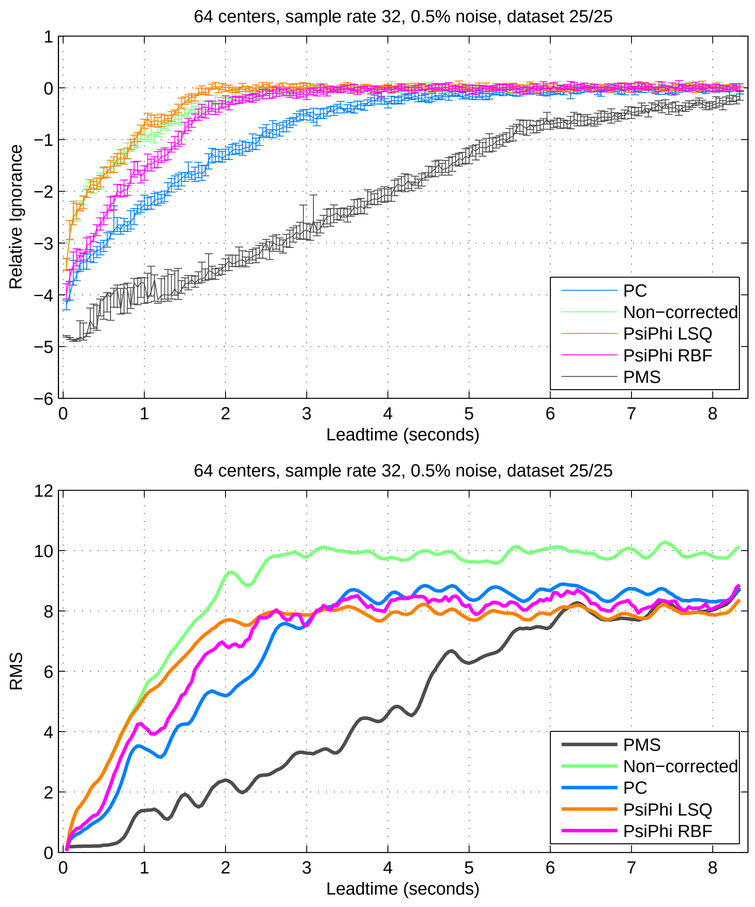

Figure 5: **Scrutinizing RMS evaluation.** Based on Ignorance the best performer of the 4 imperfect model is the PC (blue) followed by the $\Psi\Phi$ RBF (magenta) which has also significantly improved the core model and at very short leadtimes up to 2 days outperforms the PC. Both the non-corrected Φ model and the $\Psi\Phi$ LSQ quickly loose forecasting skill beyond leadtime of 2-3 days. The Ignorance of PMS (gray) is shown for comparison.

In Fig: 5 we zoom on the evaluations of the first 8 seconds of the 1200 forecasts of x variable of the Lorenz 63 system. In the top panel Ignorance is used to evaluate the model performances. In the bottom panel the RMS is used. There is a clear contradiction. RMS suggests that the $\Psi\Phi$ LSQ is significantly better then the non-corrected Φ model and potentially comparable to both the $\Psi\Phi$ RBF and the PC. The $\Psi\Phi$ LSQ forecast of the Lorenz 63, Fig: 3, demonstrates that the model is not very usefull as it quickly resorts to forecasting the mean. Despite the model's disappointing performance the RMS evaluation suggests performance comparable with the alternative models. Using RMS one may end up choosing $\Psi\Phi$ LSQ over it's competitors. Ignorance based evaluation, detects the poor performance of $\Psi\Phi$ LSQ and correctly ranks it as the worst performer.

6 Discussion

We have contrasted '$\Psi\Phi$' corrector with a new iterative approach called PC. We have shown that PC method can significantly outperform the $\Psi\Phi$ approach when considering probability forecasts. We have also argued that probability based evaluation is more useful way of evaluating nonlinear models. Using forecasts of the Lorenz 63 system we have demonstrated that scores based on least squares minimization may be potentially dangerous and lead to incorrect model selection. Although all models remain wrong, we can see more clearly which are useful.

References

[1] J. Bröcker and Smith L.A. Scoring probabilistic forecasts: The importance of being proper.

[2] K. Judd and M Small. Towards long-term predicition.

[3] E. N. Lorenz. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20:130–141, 1963.

[4] E. N. Lorenz. Irregularity: a fundamental property of the atmosphere. 36, 1984.

[5] M. S. Roulston and L. A. Smith. Evaluating probabilistic forecasts using information theory.