Climate Dynamics manuscript No.
(will be inserted by the editor)

Supplement to: Artificial neural network assisted Bayesian calibration of climate models

Tristan Hauser · Andrew Keats · Lev Tarasov

Received: date / Accepted: date

1 Analysis of error models

Here we present further analysis of the error models used for the emulators and the model discrepancy. Figure 1 shows a normal Q-Q plot of the results of scaling the differences between the mean BANN prediction and actual model output by the uncertainties estimated for each prediction using the 1σ range of the BANN posterior for that target and parameter set. These are the emulators which were used to select the final sample of Ensemble A for the perfect model experiment; i.e., they were trained on the first 80 model runs of the ensemble, which are tested against the final model runs produced for this ensemble. The fit to a Gaussian is reasonable for the most part, except for at the tails of the distribution of emulator errors, which are quite exaggerated. Our model of emulator error does not account for correlations, and so we do not expect a close fit, but it is clear that here the errors are too long tailed to be normally distributed. As such it would seem a more appropriate choice

Tristan Hauser · Andrew Keats · Lev Tarasov

Physics - Physical Oceanography

Memorial University Newfoundland, NL A1C 5S7 Canada

E-mail: lev@mun.ca

of error model would be a thicker tailed distribution. Figure 2 shows similar behaviour for the emulators which were used to select the final sample of Ensemble A for the calibration to reanalysis data. The behaviour seen here is more muted however, except for an extreme outlier.

To give an impression of the general evolution of the central tendencies of BANN performance we include Figures 3 and 4. These show the spread of RMS errors (using their mean and standard deviation) between model output and the emulator predicted values for each iteration of the calibration experiments (plots are for BANNs used to predict temperature values, and are representative of the overall BANN behaviour). Also included is the mean predicted emulator uncertainty at each iteration. It can be seen that for much of the calibration routine this value is comparable to the mean error.

To check the validity of the our model discrepancy estimates, given our poor description of the distribution of emulator predictions, we again construct Q-Q plots, this time for the differences between the outputs of the members of the final model ensemble (for the calibration to reanalysis data experiment) and the calibration targets. Each error is scaled by the associated model discrepancy and (much smaller) observational uncertainty estimate and this distribution is compared to a standard Gaussian in Figure 5. The distribution is skewed, and somewhat biased, as expected, as no attempt was made to account for bias or correlation in the model output. However, the range of the scaled errors compared to the standard Gaussian shows that the estimated model discrepancy is quite conservative. Without more sophisticated error models, we can't know what effect the over simplified emulator error model has had on the results of the model discrepancy estimates, although given the large difference in scale between the emulator and model errors it is unlikely to have been significant.

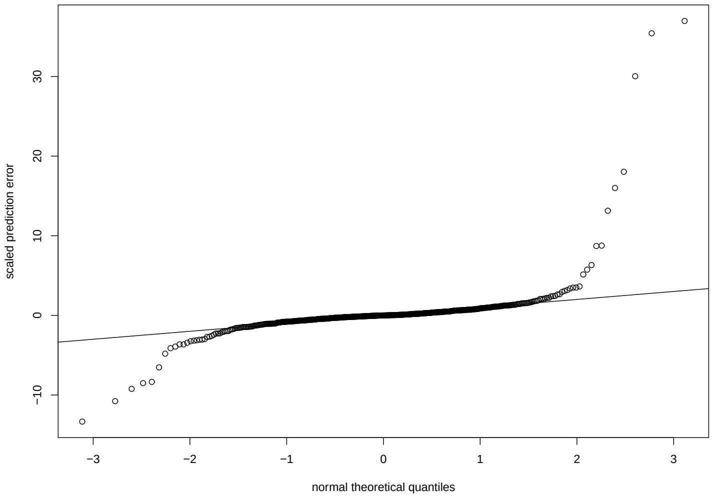

Fig. 1 Normal Q-Q plot of differences between emulator prediction and actual model output, with each error scaled by its associated emulator-predicted uncertainty. Also plotted is the line of unit slope. These are the emulators which where used to select the final sample of Ensemble A for the perfect model experiment.

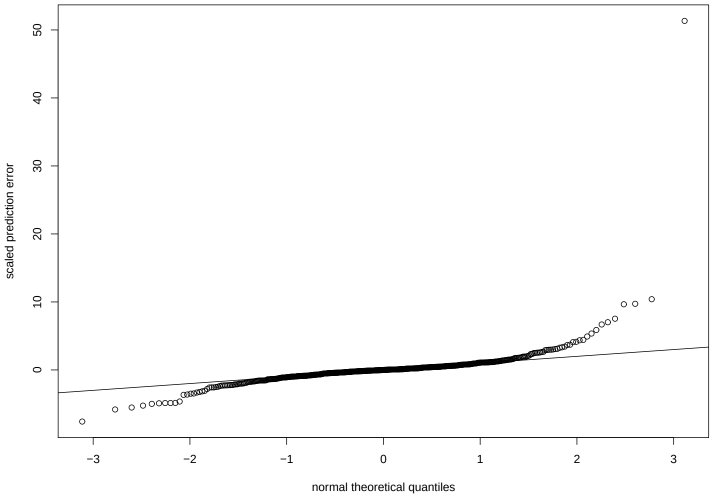

Fig. 2 Normal Q-Q plot of differences between emulator prediction and actual model output, with each error scaled by its associated emulator-predicted uncertainty. Also plotted is the line of unit slope. These are the emulators which where used to select the final sample of Ensemble A for the calibration to reanalysis experiment.

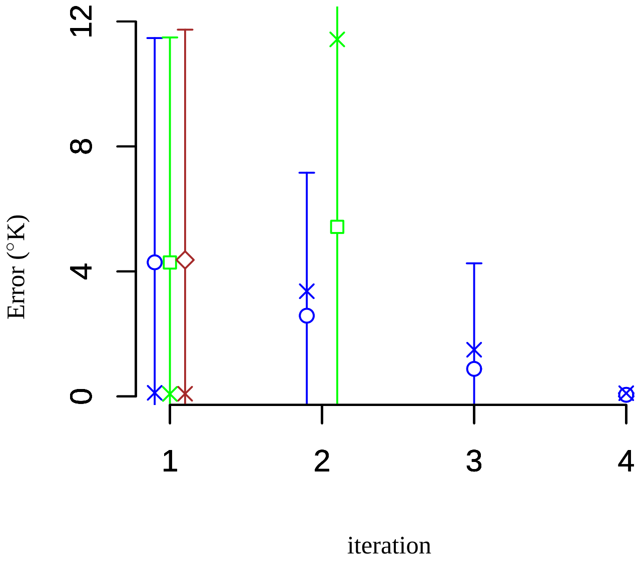

Fig. 3 Spread of RMS errors (y-axis) between actual model responses and those predicted by the emulator for each iteration (x-axis, points are offset for clarity) of the perfect model experiment, are displayed with the mean bracketed by the standard deviation. Ensembles *A*, *B*, and *C* are represented by the colours blue (circle), green (square), and brown (diamond), respectively. Crosses represent the mean predicted emulator error as estimated by the emulator.

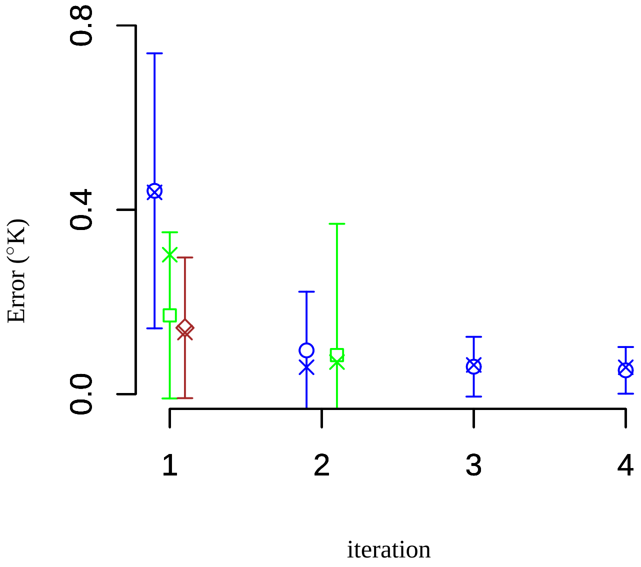

Fig. 4 Spread of RMS errors (y-axis) between actual model responses and those predicted by the emulator for each iteration (x-axis, points are offset for clarity) of the calibration to NCEP/NCAR data, are displayed with the mean bracketed by the standard deviation. Ensembles *A*, *B*, and *C* are represented by the colours blue (circle), green (square), and brown (diamond), respectively. Crosses represent the mean predicted emulator error as estimated by the emulator.

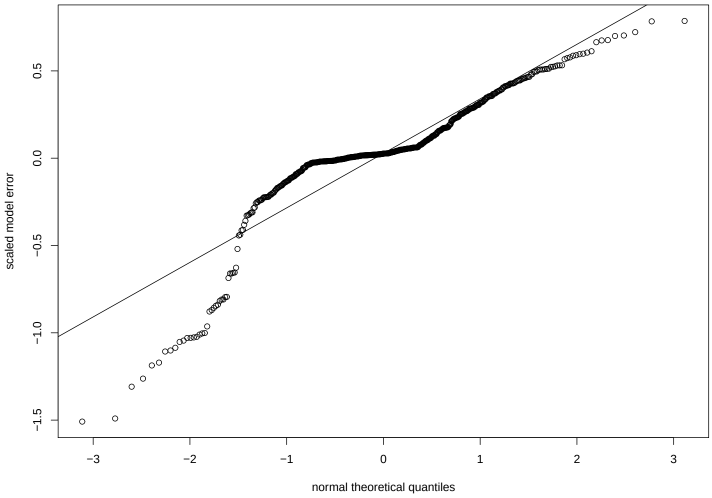

Fig. 5 Normal Q-Q plot of differences between model output and calibration targets, with each error scaled by its associated estimated model and observational uncertainty. The line of best fit to the presented points is also plotted.