Abstract

We transformed monthly ensemble temperature predictions from the NCEP Climate Forecast System (CFSv2) into probabilistic prediction using various measures of climatology and forecast uncertainty. By using information gain to evaluate the accuracy of the probabilistic predictions based on these ensembles, we were able to diagnose specific problems in the ensembles. We learned that the standard deviation of past temperatures (climatology) is a better measure of forecast uncertainty than the ensemble spread. These results led us to implement a 2-3 month lookback auto-regressive statistical model to quantify the degree to which climate dynamics in CFSv2 improve forecast skill. We found that combining climatology, recent history, and CFSv2 forecast did better than any individual component. Using the scientific Python ecosystem, we built a library for developing, evaluating, and visualizing these probabilistic forecasts. Our toolbox can be easily adapted to other ensembles.

Quantify Forecast Accuracy

The NCEP CFSv2 predicts weather. Ensembles of CFSv2 monthly hindcasts (Forecasts or $f$) cover about 30 years on an approximately 1° x 1° spatial grid.

Measure Information Gain

The information gain is the difference in the NLL of the baseline model and the model being evaluated:

$$ IG = -\log_2(\mathcal{P}(O_{lat,lon})) - \log_2(\mathcal{P}_d(O_{lat,lon})) $$

Combine Climatology and Forecasts in a New Model

The mean of our distribution was obtained using a linear auto-regressive model:

$$ \mu_{\text{fr}}(t,q) = \alpha(t,q)\mu_{\text{fr}}(t-2,q) + \beta_1(t,q)O(t-2,q) + \beta_2(t,q)O(t-3,q) $$

where $\alpha(t,q)$, $\beta_1(t,q)$, and $\beta_2(t,q)$ were fit for all times $t' < t$ at location $q$. We represented the uncertainty (which was used as the standard deviation) as the running error:

$$ \text{stdev}_{\text{fr}}(t,q) = \mu_{\text{fr}}(t,q) - O(t,q) $$

Evaluate Information Gain of Regression Based Models

We accounted for the effect of trends in the data by computing climatology as an exponentially weighted moving average (EWMA), weighing the observations in the training data using EWMA, and with a combination of both these methods.

Create Probabilistic Models

We estimated the forecast distribution ($\mathcal{P}$) using the ensemble members from a given month and year ($t$), whereas the climatology distribution ($\mathcal{P}_c$) used the climatology from that same month for all prior years; both were computed at a single latitude and longitude. We then computed the probability of the observation occurring in the distribution $\mathcal{P}(O_{lat,lon})$.

An observation where the prediction is very spread out (very uncertain) but centered around the mean may have a better RMSE than a more certain but slightly off center prediction due to RMSE’s sensitivity to relative distance from the mean.

Combine Climatology and Forecasts in a New Model

The observation was interpreted as an informational signal with a bit rate given by the terms of entropy, computed by taking the negative log likelihood (NLL) of $\mathcal{P}(O_{lat,lon})$. To compute the information gain, we first looked at the NLL of the probability of the observation occurring in the model - $\log_2(\mathcal{P}(O_{lat,lon}))$. The observation was interpreted as an informational signal with a bit rate given by the terms of entropy, computed by taking the negative log likelihood (NLL) of $\mathcal{P}(O_{lat,lon})$. To compute the information gain, we first looked at the NLL of the probability of the observation occurring in the model - $\log_2(\mathcal{P}(O_{lat,lon}))$.

The observation was interpreted as an informational signal with a bit rate given by the terms of entropy, computed by taking the negative log likelihood (NLL) of $\mathcal{P}(O_{lat,lon})$. To compute the information gain, we first looked at the NLL of the probability of the observation occurring in the model - $\log_2(\mathcal{P}(O_{lat,lon}))$. The observation was interpreted as an informational signal with a bit rate given by the terms of entropy, computed by taking the negative log likelihood (NLL) of $\mathcal{P}(O_{lat,lon})$. To compute the information gain, we first looked at the NLL of the probability of the observation occurring in the model - $\log_2(\mathcal{P}(O_{lat,lon}))$.

An observation where the prediction is very spread out (very uncertain) but centered around the mean may have a better RMSE than a more certain but slightly off center prediction due to RMSE’s sensitivity to relative distance from the mean.

An observation where the prediction is very spread out (very uncertain) but centered around the mean may have a better RMSE than a more certain but slightly off center prediction due to RMSE’s sensitivity to relative distance from the mean.

An observation where the prediction is very spread out (very uncertain) but centered around the mean may have a better RMSE than a more certain but slightly off center prediction due to RMSE’s sensitivity to relative distance from the mean.