Steps Forward

Model Comparison

To change the model comparison into a prospective comparison, we need to decide how we are going to do this. We are going to use models that depend on 1-hr behind covariates. This allows us to make one-step-ahead predictions without the mental gymnastics. This is also justifiable because current pollution concentrations are intuitively related to past (recent) weather conditions. One advantage of the missing data approach that we're abandoning is that what we do is very clear-cut, but it doesn't assess the model for its purpose. Let θ be all model parameters. Here are few model comparison approaches and my comments about them. Let me know what you think.

- **What we already did**– Select a random hold-out set and impute these held-out data using the full conditional distribution $[Y_i|\theta, Y_{-i}]$.

Here are the forward prediction approaches:

1. We could hold out data randomly again. For any time t where we want to make predictions. In this case, we can use sequential model fitting approach so that we have posterior corresponding to each t ($\pi(\theta|Y_{1:(t-1)})$). This would be a pain with our data because we have large N and would involve coding everything again. If we do this, our predictive distribution is

$$[Y_t|Y_{t'<t}] = \int_\theta [Y_t|\theta] d\pi(\theta|Y_{1:(t-1)}).$$

   In that case, we could use any subset of the data for validation without double dipping. But I'm not sure that it's computationally feasible for our problem. I could be wrong, but it'd be a pain.

2. We leave out the end of the dataset in some way (e.g. last two weeks, the last month, or the last two months) for model validation. This approach poses a couple challenges as well. This leaves us with a posterior $\pi(\theta|Y_{1:t_{\text{train}}})$, where t_{train} is the last time that we use to train our model. Now it's a question of how to forecast. Consider these three approaches:

(a) We can make h-step-ahead predictions on held-out data. For $t = t_{\text{train}} + h$, our predictive distribution is

$$[Y_t|Y_{1:t_{\text{train}}}] = \int_\theta [Y_t|\theta] d\pi(\theta|Y_{1:t_{\text{train}}}).$$

The issue that I see with this is that we are interested in one-hour-ahead forecasts, not a combination of 1,2,3,...-hour-ahead forecasts.

(b) We could treat the posterior as fixed after we train the data. Then, for $t > t_{\text{train}}$ our predictive distribution looks more like $[Y_t|Y_{1:(t-1)}, \theta]$, where θ is trained using $Y_{1:t_{\text{train}}}$. Using $[Y_t|Y_{1:(t-1)}, \theta]$ instead of $[Y_t|Y_{1:(t-1)}]$ means that our predictions don't come from the posterior predictive distribution in the way that I understand posterior predictive distributions. Do you understand what I mean?

3. We could fit the model to all of the data $Y_{1:T}$. We have a posterior that looks like $[\theta|Y_{1:T}]$. Then, for any t, we would have a predictive distribution like $[Y_t|Y_{1:(t-1)}, \theta]$. In this case, we could do validation on all of the data, but it isn't out of sample). This, of course, gives us the most information about our pollution model, but we're double dipping. To me, this approach makes sense for our final model when we are trying to assess prospective probabilities of phases and exceedance. **I'm not proposing this for model comparison, but it seems like this is probably what we'll do for our inference section. We say that it is prospective in the way that we aren't conditioning on future data, but we used future data to learn model parameters. So, we argue that it's prospective, given that we understand the process. To me, the ideal would be learning learning model parameters using 2016 data and then validate and carry out prospective inference on the 2017 data. But we aren't going to do that. So, we'll have to justify the forward-predictions using model parameters that rely on future observations.**

Regardless of the approach we choose, we will compute univariate summaries like MSE, MAE, CRPS, and coverage. Ultimately, we'll choose a model using an energy scores (ES) (a multivariate generalization of CRPS – see the citations below). For

data in $\mathbb{R}^m$, the energy score is defined as

$$\mathrm{ES}(P, \mathbf{y}) = \frac{1}{2} E_P \left\| \mathbf{Y} - \mathbf{Y}' \right\|^\beta - E_P \left\| \mathbf{Y} - \mathbf{y} \right\|^\beta, \tag{1}$$

where $\mathbf{Y}$ are predictions, $\mathbf{y}$ is an observation, and $\beta \in (0, 2)$.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102, 359378.

It is common to fix $\beta = p = 1$. **To me, it makes sense that we would use some kind of studentized or standardized observations if we are going to use this. Otherwise, the variable with the largest scale will drive our decision making. This is only part of our paper, but I think this makes our model selection stronger.**

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L. and Johnson, N. A. (2008). Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. Test, 17, 211235.

Final Model and Inference

I was arguing that we cpuld use approach 3 for inference, and I don't think that we need to make too many crazy arguments. Ideally, I think we'd use approach 1 or 2b using old data (2016) to inform our inference/predictions about 2017. It would probably eliminate our issues, but I doubt that it's worth the headache of getting all of the 2016 data from Guadalupe and Eliane. Let me know what you think.