Homework Assignment: Statistical Weather Forecast

[**Overview**](#_i7n6spv8n1tq) **1**

[**The Data**](#_20m77s600yjg) **1**

[**Assignment 1**](#_2tghspvsnjxz) **2**

[**Assignment 2**](#_xytz9empikes) **2**

[**Hints**](#_uc6kxqb8luj) **3**

[**The Rules**](#_6hmuluqnissr) **3**

# Overview

Statistical methods are proven to be powerful to post-process NWP-based forecasts and generate a corrected forecast which tries to account for the original model’s biases, or to generate a consensus forecast based on ensemble of forecasts.

In this assignment, you are about to evaluate NWP-based forecasts datasets with its corresponding observational data, and attempt **to improve** or **correct the forecasts** in a quantified way. We’ve prepared three datasets for you which includes 1 month of weather forecasts at Boston Logan International Airport (KBOS), and Hill City Municipal Airport (KHLC), with the corresponding observations.

# The Data

There are two zipped files per stations, each containing the following data:

1. Three unique NWP-forecasts named HRRR, GFS, ECMWF
   1. pt: fixed (non-changing) metadata about our location of interest (i.e. KBOS)
   2. model\_run is the cycle of the model that generated the initial forecast
   3. fcst\_time is the timestamp the forecast is valid for
   4. fcst\_dt is the number of hours lead time of the forecast (e.g. “fcst\_time - model\_run”, in hours)
2. obs: a pair of forecast parameters and their validating observations.
3. The following fields are available in this dataset (obs/model): temperature\_k (sfc\_air\_temperature), dewpoint\_k (sfc\_dewpoint), wind\_speed\_mps (u\_wind\_10m, v\_wind\_10m), sfc\_press\_pa (sfc\_pressure), cloud\_cover (cloud\_cover).

# Assignment 1

You will assess the quality of the raw NWP forecasts from the sample dataset, and propose / demonstrate some way **to improve each forecast individually**. Your deliverables (enumerated below) are the artifacts of the typical data science process which would be used for this assessment:

1. At a high-level, which forecast is generally most accurate? Present a visualization or table which summarizes the accuracy of the raw forecast data using standard statistics of your choice
2. **How might you statistically correct each forecast?** Design and document an experiment to create a bias correction scheme for the forecast, leveraging the available forecast data and corresponding observations.
   1. Justify the design of your experiment and scheme.
3. Do your statistically-corrected forecasts improve on the raw forecasts?
   1. Create a summary visualization graphic or table with statistics facilitating a comparison between the two raw and corrected forecasts.
   2. What is an overall improvement using this correction?
4. How might you try to improve upon this method in the future? Suggest a method or technique that you might consider implementing to solve this statistical forecasting problem, assuming that you had 3 months to develop and demonstrate its viability.

# Assignment 2

You will assess the quality of the raw NWP forecasts from the sample dataset, and propose / demonstrate some way **to generate a consensus forecast**. Your deliverables (enumerated below) are the artifacts of the typical data science process which would be used for this assessment:

1. At a high-level, which forecast is generally most accurate? Present a visualization or table which summarizes the accuracy of the raw forecast data using standard statistics of your choice
2. **How might you generate a consensus (i.e. single, most accurate) forecast using these ensembles?** Design and document an experiment to generate a consensus forecast, leveraging the available forecast data and corresponding observations.
   1. Justify the design of your experiment and scheme.
3. Do your consensus forecasts improve on the raw forecasts?
   1. Create a summary visualization graphic or table with statistics facilitating a comparison between the two separate forecasts.
4. How might you try to improve upon this method in the future? Suggest a method or technique that you might consider implementing to solve this statistical forecasting problem, assuming that you had 3 months to develop and demonstrate its viability.

# Hints

* If you choose to train a statistical model, how will you prevent data leakage from your training to validation/evaluation?
* You have multiple overlapping timeseries of forecasts for the majority of your individual forecast timestamps in the dataset.
* In this exercise, you are welcome to apply the analysis first to temperature, then wind speed and, and finally to dewpoint temperature.

# The Rules

1. **Choose only 1 assignment** based on your preference.
2. The analysis should be developed in Python and presented in a literate coding framework (i.e. Jupyter Notebook) with proper comments, documentation, and organization.
   1. Collection of scripts are acceptable with proper documentation that reviewers can follow and reproduce.
3. Submit your slide deck alongside the analysis code when you return this assignment.
4. **You are not required to spend more than 4-5 hours on this assignment.**
5. We encourage leveraging the scientific Python ecosystem of tools that could enhance your work.
6. Submit your answers with any supporting materials (documentation of methodology, sample files, namelist, presentation, graphs etc) within 7 days of receiving the assignment.
   1. If you run into any roadblocks, feel free to review online resources such as Stack Overflow, or reach out to us directly.

Good luck!

Tomorrow.io team

Last Revision: July 2021