Climate Data Challenge

Lawrence Livermore National Laboratory
Data Heroes Program

July 14–August 7, 2015

1 Introduction

The third data challenge for the Data Heroes Program is a multi-class classification problem involving three international climate models. Climate forecasting involves complex physics-based models that characterize the earth as a physical system. Different climate models involve alternative ways to characterize this physical system, resulting in differing forecasts. As all model outputs differ and cannot all be correct, there is inherent bias in the model forecasts.

This competition is not focused on judging accuracy of any model, but rather whether the model biases are dominant enough to make differentiation of location forecasts possible. This competition uses the historical simulations for near-surface air temperature (model output *tas*) as the basis for model differentiation.

2 Climate models

The historical simulations start with initialization in January 1850 and run on a monthly time-step until December 2005. We throw out the first fifty years of data as an initialization or "burn-in" period and start the competition data in January 1900.

There are three prominent international climate models generating the data for the contest. We label the models *Model 1*, *Model 2*, and *Model 3*, as to not give away the source of the data. Five runs of each model comprise the competition data.

3 Training data

There are three data files representing the data of each model, these files are titled *model#_train.csv*. The files contain complete data for four runs (of the five) at each location (unique latitude and longitude combination). The other run is withheld from the set of five runs at random. Note that the run selected is random at each location. The order of the four runs presented in the data for each location has been scrambled, so you don't know if the first run is listed first or fourth (or not at all if that is the run selected for withholding at that location).

The first three columns of each data set are an identification index, location latitude (in degrees North), and location longitude (in degrees East). These columns have headings `""`, `"V1"`, and `"V2"` The remaining columns (starting with heading `"V3"`, ending with heading `"V1274"`) are the

model near-surface air temperature output (in Kelvin) starting January 1900 (ideally mid-month observations), ending in December 2005, for a total of 1272 monthly values.

The model grid information and data set sizes are listed in the following table:

	Latitude divisions (start/stop/step)	Longitude divisions (start/stop/step)
Model 1	192 (−90 / 90 / 0.94241)	288 (0 / 358.75 / 1.25)
Model 2	128 (−88.92774 / 88.92774 / 1.40044)	256 (0 / 358.59375 / 1.40625)
Model 3	160 (−89.14152 / 89.14152 / 1.12128)	320 (0 / 358.875 / 1.125)

	Observations	Columns	File Size
Model 1	221184	1275	2.51 GB
Model 2	131072	1275	1.49 GB
Model 3	204800	1275	2.32 GB

4 Test data

The object of the contest is to correctly classify the generating model for the 30000 observations of the test data in file *model_test.csv*. These observations contain latitude and longitude information for locations, rounded to the nearest 2.5 degrees. To be precise, if x represents a location's latitude and longitude, the rounded value in the test data is equal to $2.5 \cdot \text{round}((x + \epsilon)/2.5)$, where ϵ is a very small non-zero positive value to prevent ambiguity when rounding a half.

There were 55296 location observations withheld in the model 1 data, 51200 observations withheld from model 2, and 32768 observations withheld from model 3, as there is one run from each location of the grid withheld. From each of these withheld sets, 10000 are randomly selected for the test set. So there are 10000 observations each from the three models in the test set to be classified. The observations are ordered by location latitude and longitude, noting that observations within any latitude and longitude combination have been shuffled.

The file *template.csv* is a sample submission template with two columns, *id* relating to the row identifier of the test data and *model* a user-entered value from $\{1, 2, 3\}$ representing the choice of generating model for that observation.

5 Submitting classification results

Submission text/csv files fitting the format specified in *template.csv* will be judged by the Commissioner. Submit your file with the lone word "submission" (all lower case) in the subject line to `llnl.commish@gmail.com`. Attach your classification file with the name format {*team-name*}_{*submission-number*}.csv. The classification score will be reported to teams according to the submission number—the score being the correct classification proportion for a random selection of 25% of test observations. We will set a limit of five submissions per weekday, resetting each morning. Submissions not matching the above format will not be judged. At the end of the competition, each team will select a single submission as their final results to be scored to the other 75% of observations. The proportion of correctly classified observations for this private 75% of the test set will be the team's final score in the competition.