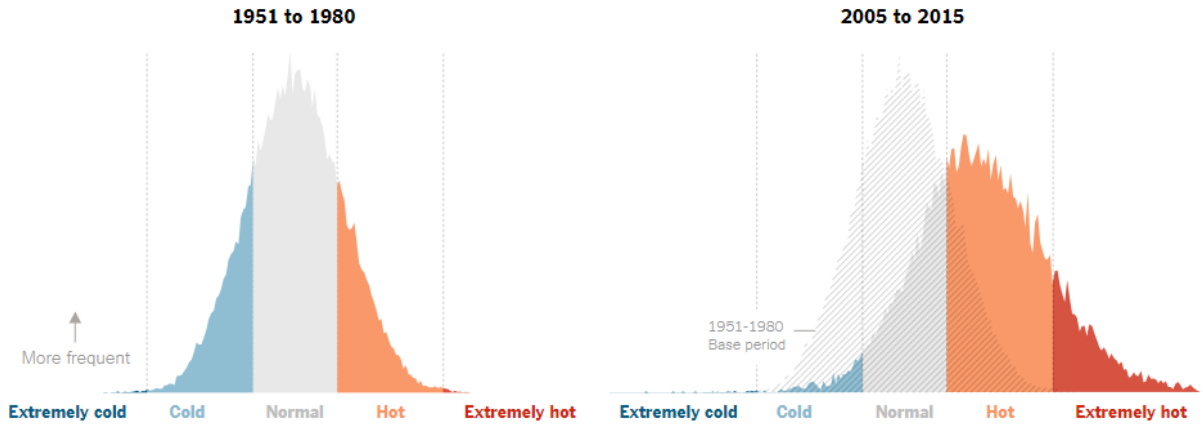
Statistical modelling and pattern recognition for predicting evolution of temperature forecasts

Company Background

BP is a global energy company that is leading the transition to a lower carbon future. No matter your role here, we each play a part in progressing the way we deliver heat, light and mobility to the world. Our business is the exploration, production, refining, trading and distribution of energy. BP operates with business activities and customers in more than 80 countries across six continents and every day we serve millions of customers around the world. Integrated Supply & Trading (IST) is BP's commercial face to the global energy and commodity markets. We market the company's upstream production of hydrocarbons, secure feedstocks for our refinery system and provide services to external customers, including fuel supply and hedging solutions.

Problem Background

Global energy companies deeply care about the weather in order to forecast energy demand for most of Europe and North America. In the US, the energy demand peaks twice a year driven by heating demand in winter months and power demand in summer months. With extreme weather becoming the norm especially during summer months (see the chart below that depicts how summer months have become more extreme globally), energy companies such as BP need to pay a close attention to not only the current forecast of temperatures but also to how the forecast may evolve in time (known as the forecast of forecast).



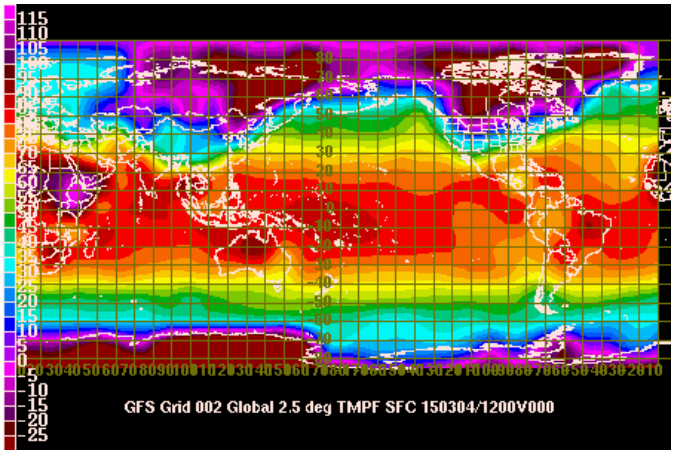
Source: NY Times

Currently there are two global numerical weather prediction (NWP) systems relied upon for temperature forecasts by academics, industry, enthusiasts and the governments across the world. These are –

1. Global Forecast System (GFS) – Run by the US National Weather Service, this model runs 4 times a day and produces forecasts for up to 16 days in advance at varying spatial resolution (13 km to 27 km). This model is a basis for derivative models such as Global Ensemble Forecast System (GEFS) and North American Ensemble Forecast System (NAEFS).
2. ECMWF European Model – Run by European Centre for Medium-Range Weather Forecasts (ECMWF), this model also runs 4 times a day. Like GFS, there are derivative models run by the agency with varying resolution to balance computational cost and accuracy.

Data

The numerical models (GFS or ECMWF) divide the globe into a grid and predict temperatures for the grid points at regular time increments out the next 15 days. Shown below is a snapshot of the output from GFS model.



For our use cases, temperature data is condensed into Average Daily Temperature representing a single value for each location and for each day.

For the scope of this project, you will be provided with 15-day temperature forecast for the period of 1-Mar-2007 to 28-Feb-2019 for 150 locations that fall within the continental US. This dataset is sourced from a vendor that uses proprietary algorithms to average and bias-correct the intraday outputs of GFS and ECMWF weather runs along with additional human forecaster nudging at times, resulting in a single forecast vector for each day. A sample from the dataset is shown below.

RUN_DATETIME	LOCATION_CODE	OBSERVATION_DATE	TEMPERATURE
3/1/2007	1	3/1/2007	54.5
3/1/2007	1	3/2/2007	47
3/1/2007	1	3/3/2007	40.5
3/1/2007	1	3/4/2007	37
3/1/2007	1	3/5/2007	40.5
3/1/2007	1	3/6/2007	44.5
3/1/2007	1	3/7/2007	48.5
3/1/2007	1	3/8/2007	51.5
3/1/2007	1	3/9/2007	54.5
3/1/2007	1	3/10/2007	52
3/1/2007	1	3/11/2007	49.5
3/1/2007	1	3/12/2007	51.5
3/1/2007	1	3/13/2007	54.5
3/1/2007	1	3/14/2007	54.5
3/1/2007	1	3/15/2007	52.5
3/2/2007	1	3/2/2007	47.5
3/2/2007	1	3/3/2007	42.5
3/2/2007	1	3/4/2007	37
3/2/2007	1	3/5/2007	41.5
3/2/2007	1	3/6/2007	46
3/2/2007	1	3/7/2007	50.5
3/2/2007	1	3/8/2007	53
3/2/2007	1	3/9/2007	53.5
3/2/2007	1	3/10/2007	51.5
3/2/2007	1	3/11/2007	52.5
3/2/2007	1	3/12/2007	53.5
3/2/2007	1	3/13/2007	56.5
3/2/2007	1	3/14/2007	54.5
3/2/2007	1	3/15/2007	52.5
3/2/2007	1	3/16/2007	50.5

1. RUN_DATETIME: The ‘as of date’ for the forecast
2. LOCATION_CD: Location code
3. OBSERVATION_DATE: The forecast target date
4. TEMPERATURE: Average Daily Temperature for the location

Problem Description

Each day's forecast at a given point is comprised of 15 numbers which represent temperatures for days 0-14 in the future. As we move to the next day the forecast vector rolls by 1 day such that tenor 14 of previous day's forecast now becomes tenor 13 of current day's forecast. Predicting the weather forecast for the next day is to predict the 15-dimensional vector for the next day. The table shown below illustrates the structure of the data.

	TEMPERATURE																				
OBSERVATION_DATE	3/1/2007	3/2/2007	3/3/2007	3/4/2007	3/5/2007	3/6/2007	3/7/2007	3/8/2007	3/9/2007	3/10/2007	3/11/2007	3/12/2007	3/13/2007	3/14/2007	3/15/2007	3/16/2007	3/17/2007	3/18/2007	3/19/2007	3/20/2007
RUN_DATETIME																				
3/1/2007	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x					
3/2/2007		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x				
3/3/2007			x	x	x	x	x	x	x	x	x	x	x	x	x	x	x			
3/4/2007				x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		
3/5/2007					x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
3/6/2007						x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Problem Statement

The problem we present to you is as follows: estimate, for each time point t, the conditional distribution of the 150 × 15-dimensional vector corresponding to the 15-day forecast at each location, given the historical data up to time t-1.

Note: datasets are provided to the working group under a Confidentiality Agreement signed on behalf of the University of Cambridge.