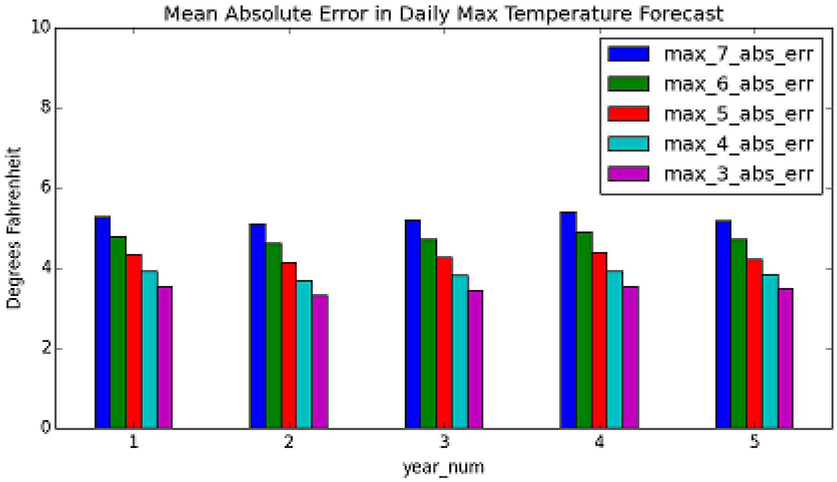
Accuracy of NWS Medium Range Weather Forecasts

Jan 8, 2011 – Jan 7, 2016



General Assembly – Data Science class project

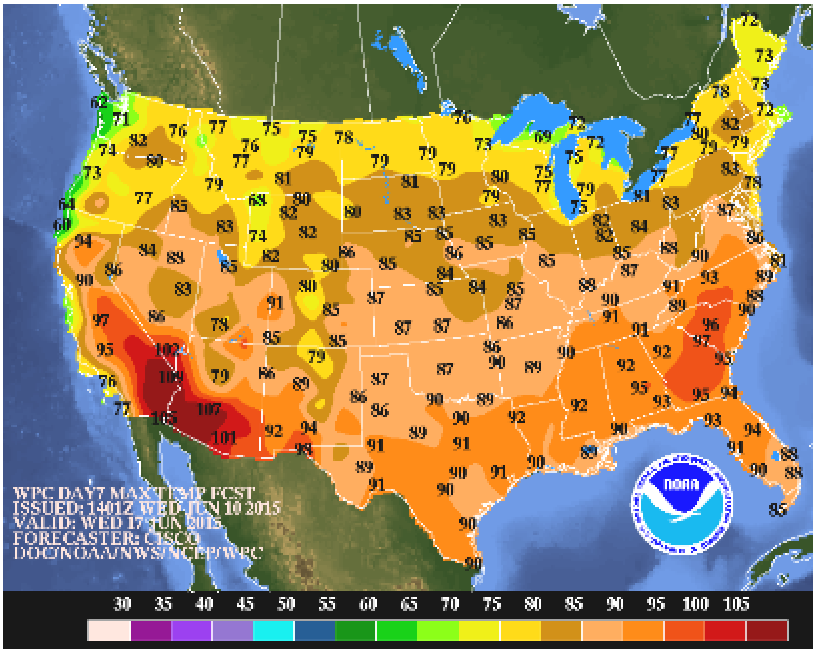
Bruce Aker

January 21, 2016

Hypothesis: Errors in weather forecasting are decreasing and there is very little systematic error

I. Forecast Data: Historical weather forecasts are available from WPC* (part of NWS) in graphical format. These are medium range forecasts (valid for 3, 4, 5, 6 and 7 days ahead) and include daily maximum and minimum temperature and the probability of precipitation for various weather stations around the continental United States.

Maximum temperature forecast issued 6/10/15 valid for 6/17/15 (7 day forecast)



DAY7_MAX_2015061012_filled.gif

*See Addendum for abbreviations

II. Obtaining the forecast weather numbers: the forecast numbers are embedded in the graphic image (.gif) of a map of the U.S. A procedure was developed to extract these numbers

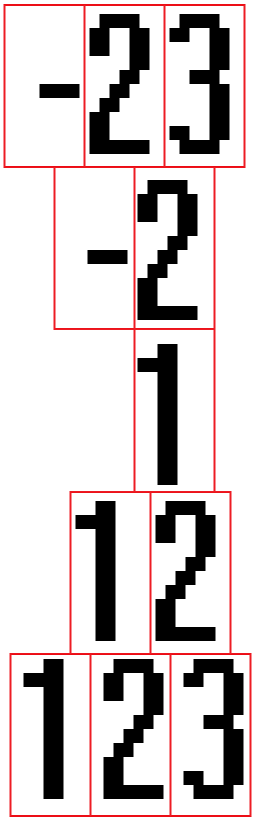
III. Extraction Procedure: the numbers contained up to three characters (within range -99 to 999) from a set of eleven characters (0, 1, 2,…, 9, -). The character font, color (black), size, location was (almost) always the same. There was no other black color, no anti-aliasing and the character overwrote the other colors. The strategy was to search the pixel array at certain locations for the pixel patterns of the characters comprising the number using Python (poor man's OCR).

Python tuple with 1's (pixel in the character) highlighted

```
((1,1,1,1,1,1,1,1,0), # 2 upside down
 (1,1,1,1,1,1,1,1,0),
 (0,1,1,0,0,0,1,1,0),
 (0,0,1,1,0,0,0,1,0),
 (0,0,0,1,1,0,0,0,0),
 (0,0,0,0,1,1,0,0,0),
 (0,0,0,0,0,1,1,0,0),
 (0,0,0,0,0,1,1,1,0),
 (0,0,0,0,0,1,1,1,0),
 (1,0,0,0,0,1,1,1,0),
 (1,1,0,0,1,1,1,1,0),
 (0,1,1,1,1,1,1,0,0),
 (0,0,1,1,1,1,0,0,0)),
```

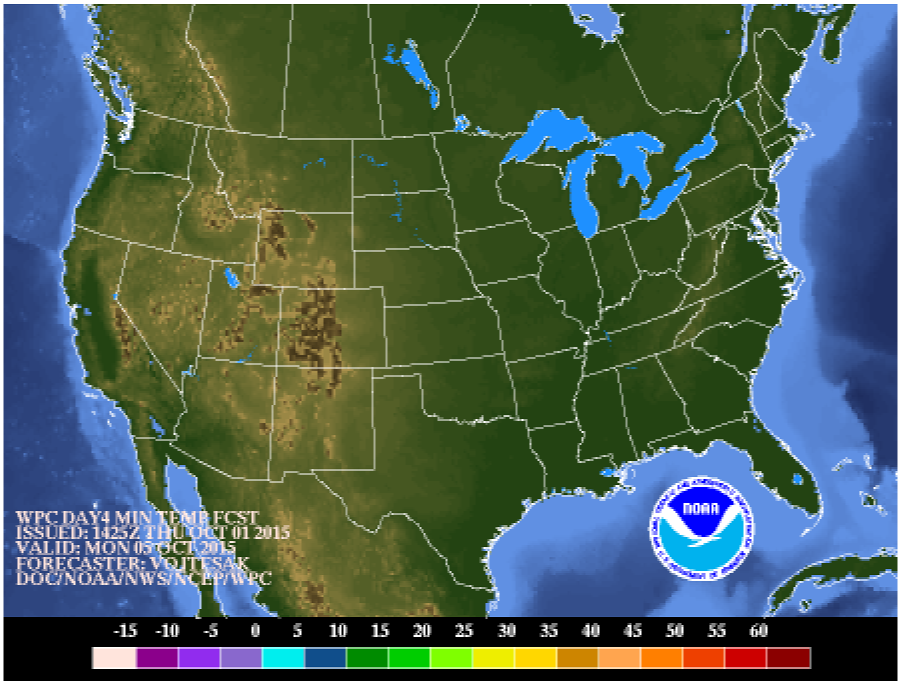
Searched for 88 weather stations, identified by ICAO ID, e.g. KSEA for Seattle Tacoma airport

The starting position of the number varied depending upon number of characters and negative or positive.



Search for negative numbers first, otherwise might erroneously find a positive number

IV. Missing Data: Some of the graphics were missing numbers:



2015100112_DAY4_MIN_filled.gif

- Some image files were completely missing: 2/29/12, 3/3/13
- One image file was missing it’s number for Muncie, IN
- Some early image files missing numbers for several locations
- Image size smaller (and smaller font) for before ~December 2012
- Precipitation forecast mostly not available in 2011

V. Obtaining GIF image files: systematically named and publicly available via HTTP.

www.wpc.ncep.noaa.gov/archives/medr/20150125/DAY7_MIN_2015012512_filled.gif

Downloaded with Python using the requests module and then written to disk as a binary file

VI. Convert .GIF images to .BMP images:

• GIF is a compressed format – harder to work with the pixel data

• BMP is uncompressed format – pixel data laid out systematically in memory

• Used Python PIL module (Python Image Library) to convert

VII. Obstacles overcome:

1.) Subtle file name change on 12/8/12:

DAY7_MIN_20150125_filled.gif DAY7_MIN_2015012512_filled.gif

2.) Some numbers were to the right or left by one pixel, for example, a few locations display single digit positive numbers one pixel to the left in the precipitation forecast images. Two other cases like this.

VIII. Obtaining the observed weather data:

•Data is available from NOAA > NCEI (formerly NCDC) > CDO > Web Services v2

•Used Python requests and JSON modules

•Obtained for each day data types: TMax, TMin, PRCP, SNOW

www.ncdc.noaa.gov/cdo-web

•All responses are JSON and are a single item or a collection of items with metadata

•Limited to 5 requests per second and 1000 requests per day

•JSON response row limit (in 'results' list) defaults to 25, max is 1000 (set 'limit' query string)

•Need a token, easily obtained via email

Obstacles overcome:

1.) Find set of weather stations in which both forecast and observables was available

2.) Find weather station identifier needed by CDO: WBAN ID, not COOP ID as specified in examples on website, not ICAO ID (cross-reference the IDs using the MSHR)

IX. Accuracy of data:

•Several forecast numbers extracted from the image files were compared to the numbers displayed in the images and found to be identical.

•Several of the downloaded observable numbers were compared to another NWS website and found to be identical

X. Combining the forecast and observed data:

Long table of Forecast data:

```
wea_stn_cd,issue_date,forecast_type_day,wea_num,valid_date_calcd
KABE,20110101,min_7,20,20110108
KABQ,20110101,min_7,25,20110108
KACV,20110101,min_7,38,20110108
KACY,20110101,min_7,20,20110108
KALB,20110101,min_7,17,20110108
.
.
.
2.66 million rows, 86 MB
```

Long table of observations:

```
wea_stn_cd,valid_date,wea_num_type,wea_num
KABE,20110108,prcp,0.039
KABE,20110108,tmin,9.0
KABE,20110108,tmax,27.0
KABE,20110108,snow,0.59
KABE,20110109,prcp,0.0
KABE,20110109,tmin,21.0
KABE,20110109,tmax,30.0
KABE,20110109,snow,0.0
.
.
.
626,000 rows, 15 MB
```

So, using Python and pandas, pivot on forecast_type_day and wea_num_type (index = Wea_Stn, Valid_date) and join to get…

…one wide table with all data:

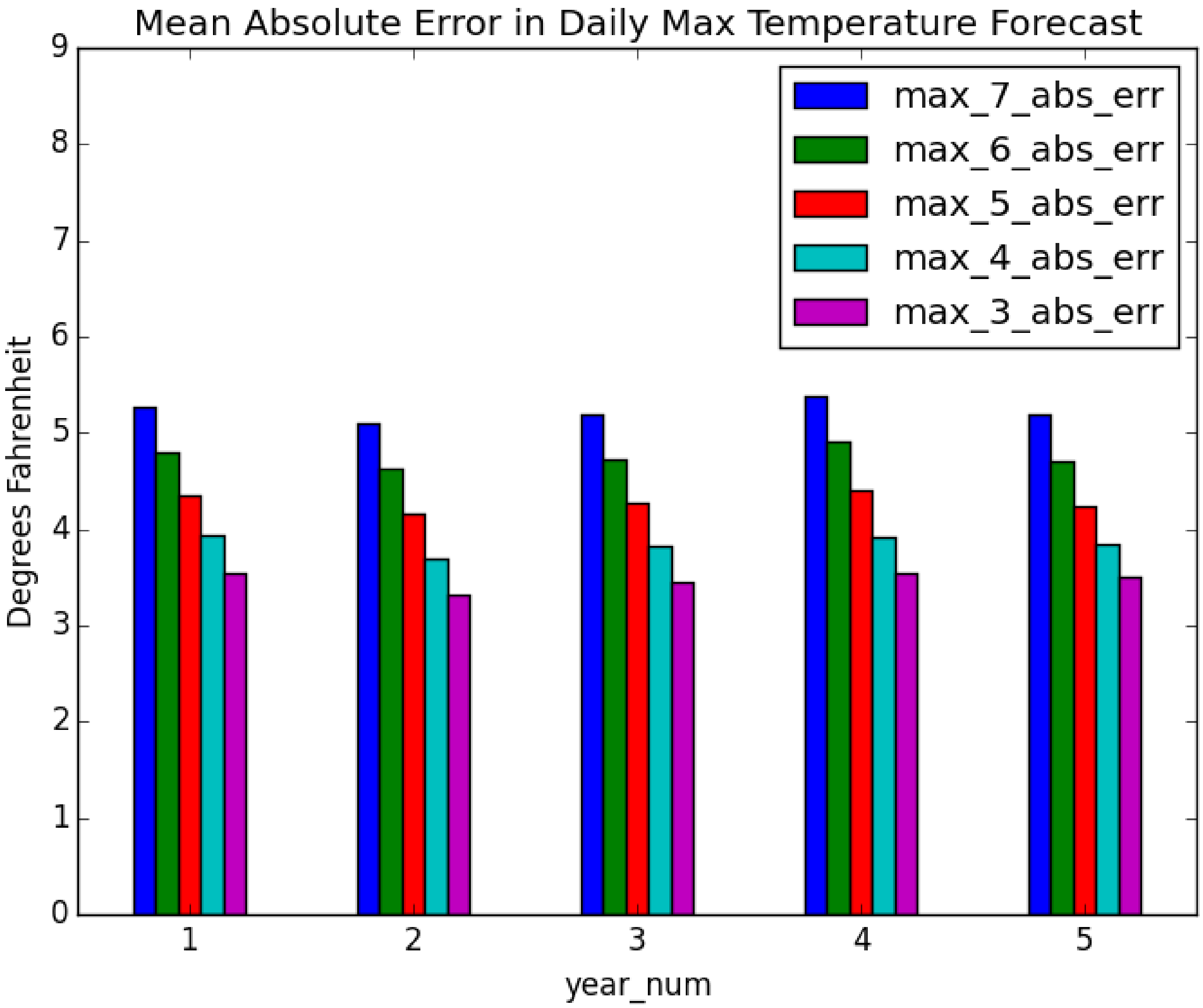
```
wea_stn_cd,valid_date,max_3,max_4,max_5,max_6,max_7,min_3,min_4,min_5,
min_6,min_7,pop1_3,pop1_4,pop1_5,pop1_6,pop1_7,pop2_3,pop2_4,pop2_5,
pop2_6,pop2_7,prcp,snow,tmax,tmin

KGRB,20151224,36.0,38.0,39.0,40.0,42.0,29.0,32.0,32.0,31.0,32.0,40.0,
52.0,38.0,31.0,27.0,17.0,23.0,19.0,16.0,23.0,0.012,0.0,37.0,30.2
.
.
.
160,000 rows, 18 MB
```

Unique row ID: wea_stn_cd (e.g. KSEA) and valid_date (YYYYMMDD)

XI. Analysis:

Calculate average of Abs(Forecast minus Observed) and plot by year_num

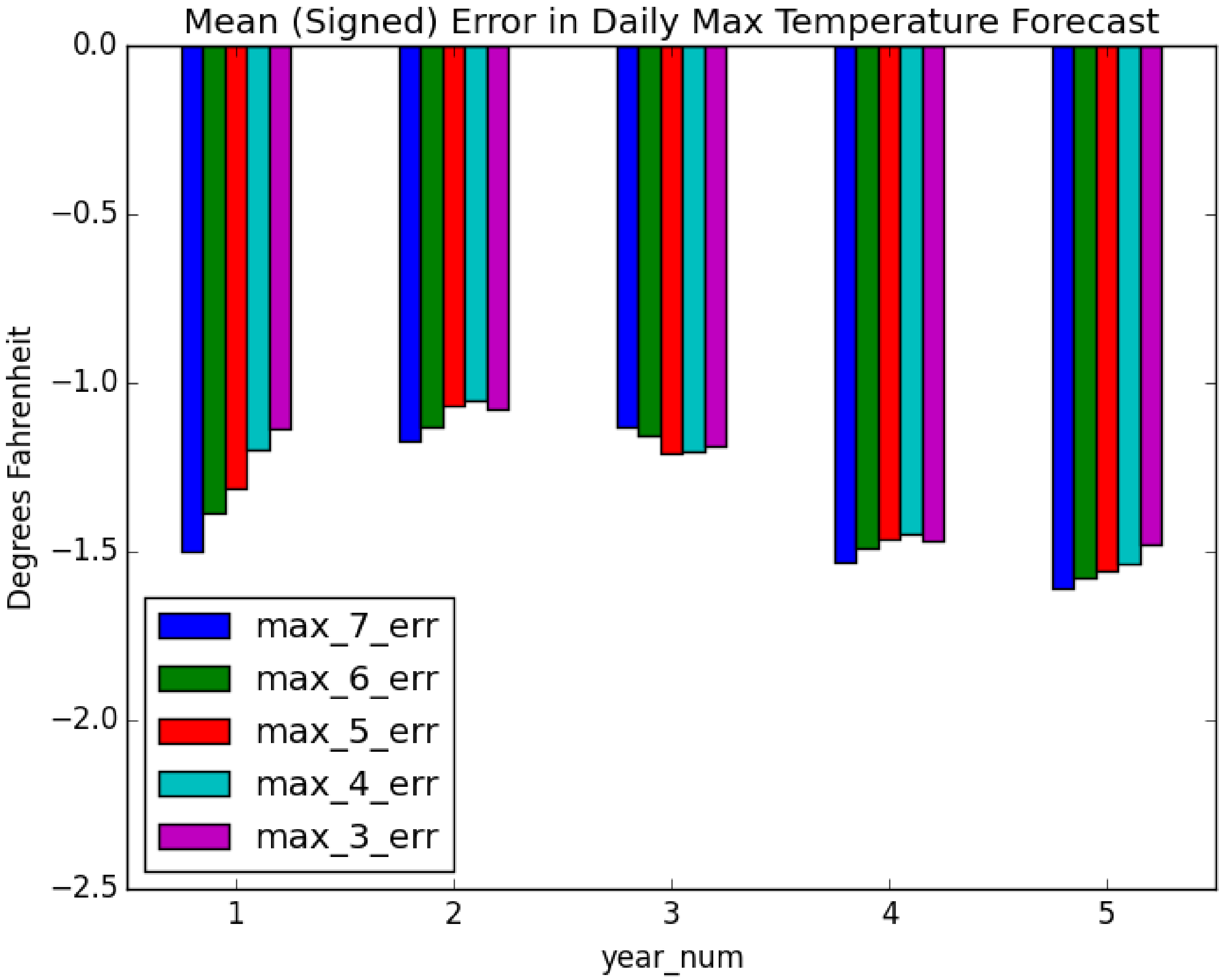
(year_num = 1 is 1/8/11 – 1/7/12)



Not a long enough time span (5 years) to see if forecast accuracy is improving. Can see that, for example, the 3 day forecast is more accurate than the 7 day forecast.

Calculate average of Forecast minus Observed and plot by year_num

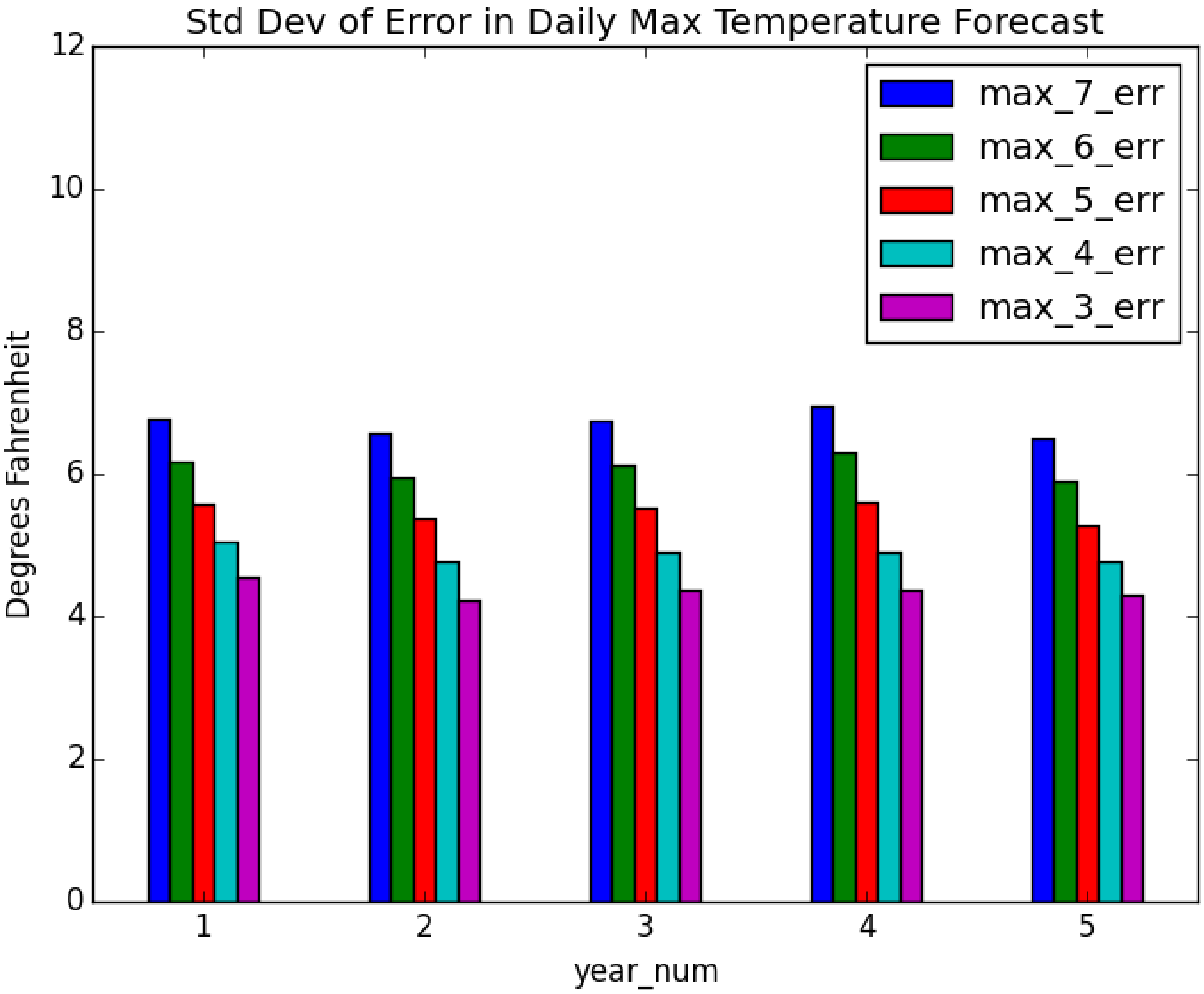
(systematic error, negative number means forecast too low)



Forecast for daily maximum temperature are systematically too low.

Calculate standard deviation of Forecast minus Observed and plot by year_num

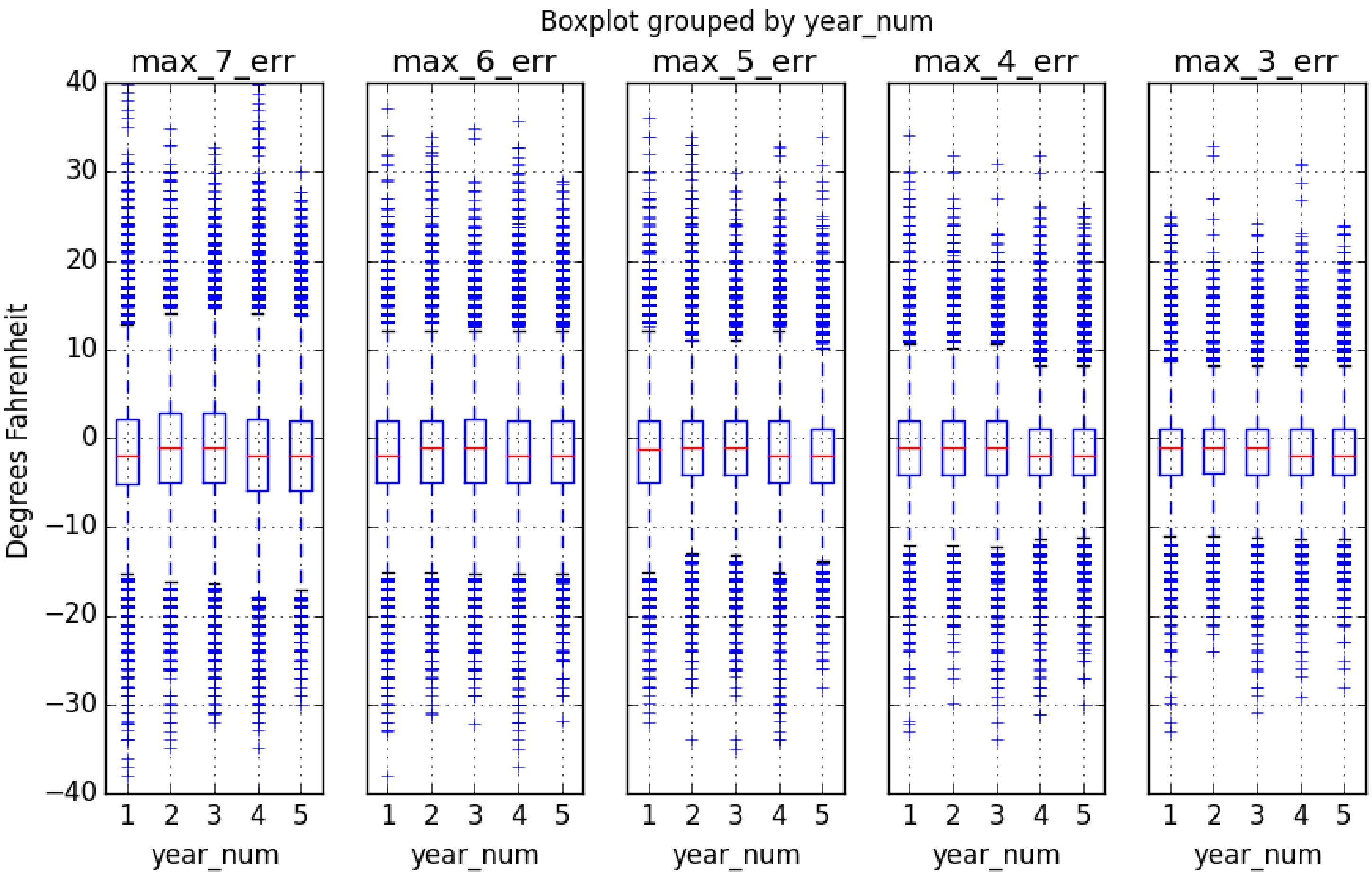
(random error)



3 day forecast is more accurate than the 7 day forecast.

Box Plot(s) summarizes the data best

(shows location and variation, medians all below zero)

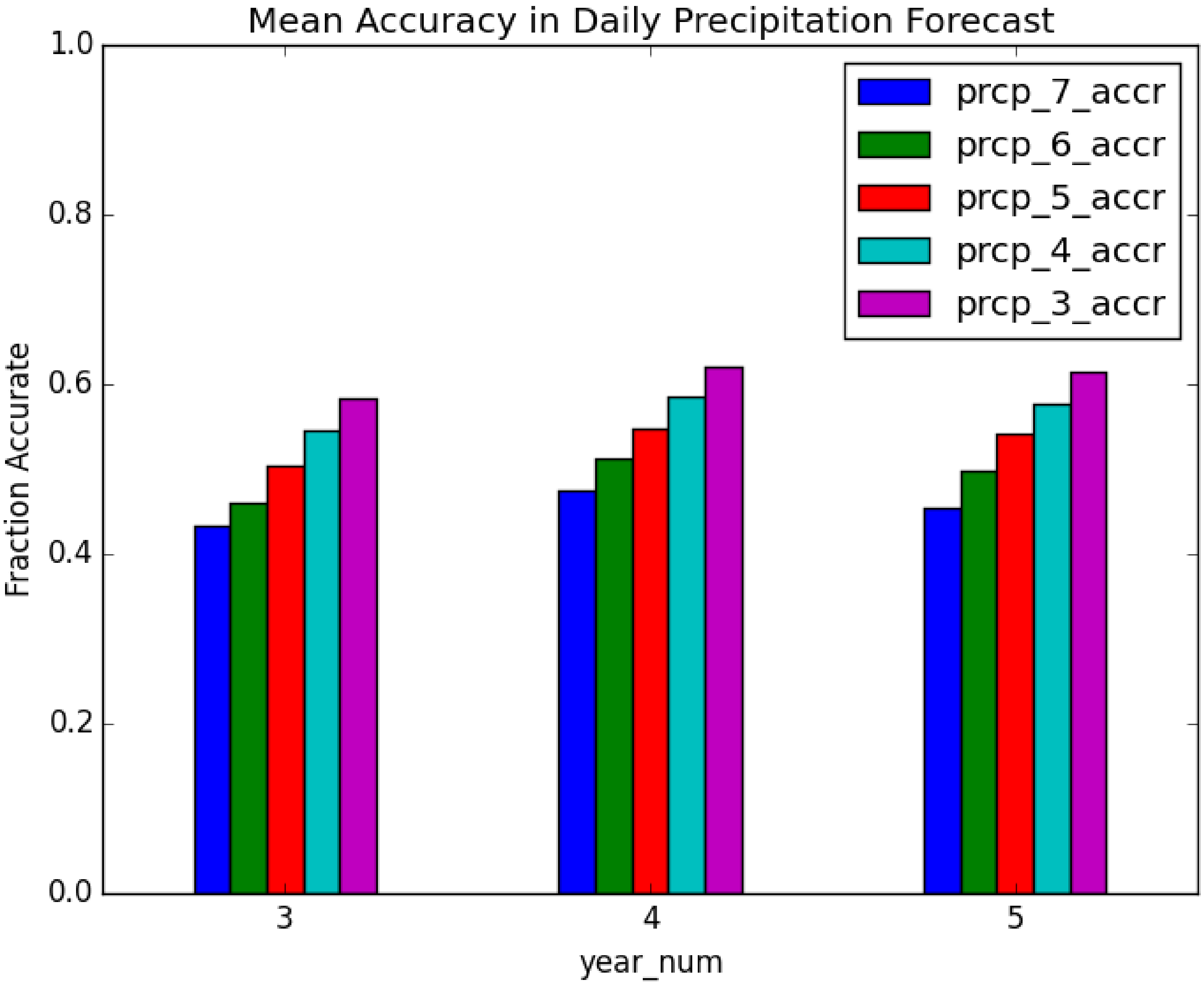


Median error is consistently about -1.0 to -1.5°F

If Probability of Precipitation <= 5% and no rain or snow then score 1

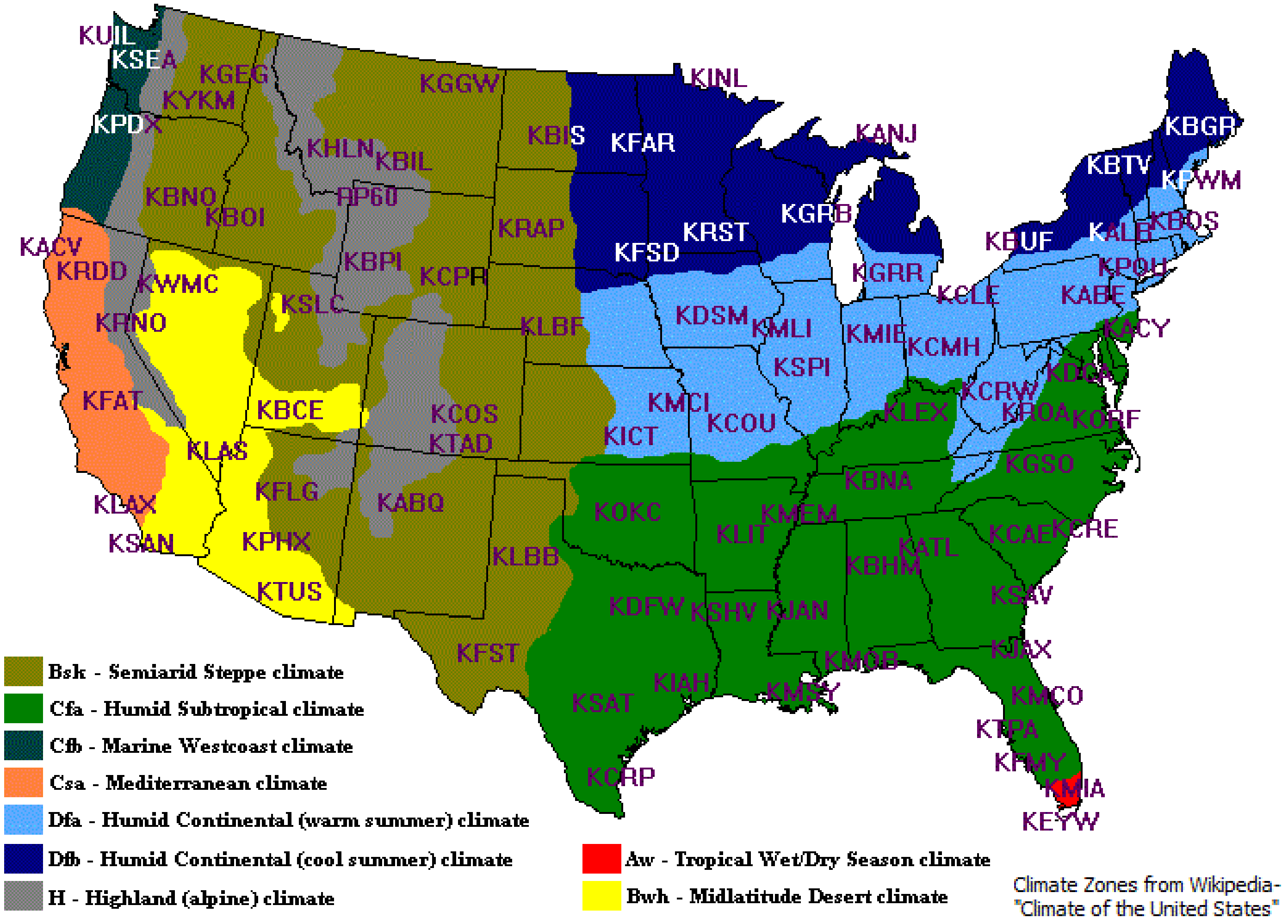
Or if Probability of Precipitation > 5% and measureable rain or snow then score 1

Otherwise score 0



XII. To do: analyze errors by climate zone

Weather Stations and Climate Zones of the Continental United States



Climate Zones from Wikipedia-
"Climate of the United States"

XIII. More To Do:

1.Investigate outliers

2.Investigate missing data

3.Check for duplicate data

4.Check for non-ASCII characters

5.Analyze errors by season, month of the year, etc.

6.Analyze errors by variation of daily normal values (e.g. variation of normal maximum temperature for a given day of the year and weather station)

7.Obtain forecast data from the nonstandard images, nonstandard font

8.Further investigate accuracy of data

ADDENDUM

All units either °F, inches, or percent (for pop1, pop2, dly-prcp-pctall, dly-snow-pctall)

GENERAL

COOPID - Weather station ID, 6 digits, sometimes prefaced with USC00
Error = forecasted – observed (for temperatures)
ICAO Code/ID - Identifier for weather station (e.g. KSEA for Seattle Tacoma airport)
STNID - (weather) station ID, issued by NCEI (formerly NCDC), 8 digits
stn_id_cdo - (weather) station ID needed by CDO website (WBAN, COOP,...), prefaced with a key, e.g. GHCND:USW00024243, COOP:USC00123456
WBANID - Weather station ID, 5 digits, sometimes prefaced with USW000
Wea_stn_cd – ICAO weather station code/ID (e.g. KSEA for Seattle-Tacoma airport)
Weather number – value for tmax, tmin, prcp, snow, predicted min/max temp, POP
Z = UTC, e.g. 1425Z = 14:25 UTC (Universal Coordinated Time)

ABBREVIATIONS

CDO - Climate Data Online
COOP - Cooperative Observer Program
GHCN - Global Historical Climatology Network
HPC Hydrometeorological Prediction Center (now WPC)
ICAO - International Civil Aviation Organization
MSHR - Master Station History Report
NCDC - National Climate Data Center (now NCEI)
NCEI - National Centers for Environmental Information (née NCDC)
NCEP National Centers for Environmental Prediction
NOAA - National Oceanic and Atmospheric Administration
NWS - National Weather Service
OCR – Optical Character Recognition
WBAN - Weather Bureau, Air Force, Navy
WFO - Weather Forecast Office
WPC - Weather Prediction Center (née HPC)

ADDENDUM CONTINUED:

FORECASTS

Forecast_Day – 3 thru 7 days ahead (WPC medium range forecasts)
Forecast_Type – Min/max temp or POP
Issue date – date forecast was issued (YYYYMMDD)
MAX_3 – predicted maximum temperature 3 days ahead (resolution 1°F)
MAX_7 – predicted maximum temperature 7 days ahead (resolution 1°F)
MIN_3 – predicted minimum temperature 3 days ahead (resolution 1°F)
MIN_7 – predicted minimum temperature 7 days ahead (resolution 1°F)
POP1 – Probability of Precipitation (>= 0.01") at 1200Z (percent)
POP2 – Probability of Precipitation (>= 0.01") at 0000Z (percent)
Valid date – date forecast is valid for (YYYYMMDD)
Valid_date_calcd – calculated from issue date and forecast day (e.g. 3,4,5,6,7) (YYYYMMDD)
wea_num - value for forecasted min/max temp, POP

OBSERVABLES – actual weather data

prcp - observed rain amount, does not include snow, resolution ~0.004 inches
snow - observed snowfall amount, resolution ~0.04 inches
tmax – observed max temp, resolution ~0.2°F
tmin – observed min temp, resolution ~0.2°F
Valid date – date observation was made (YYYYMMDD)
wea_num_type - tmax, tmin, prcp, or snow
wea_num - value for tmax, tmin, prcp, snow

DAILY NORMALS (calculated from 1981 - 2010; resolutions: temperature 0.1°F, precipitation 0.01", snowfall 0.1", probability 0.1%)

(dly-)tmax-normal - Long-term averages of daily maximum temperature
(dly-)tmax-stddev - Long-term standard deviations of daily maximum temperature
(dly-)tmin-normal - Long-term averages of daily minimum temperature
(dly-)tmin-stddev - Long-term standard deviations of daily minimum temperature
(dly-)prcp-50pctl - 50th percentiles of daily nonzero precipitation totals for 29-day windows centered on each day of the year
(dly-)prcp-pctall - Probability of precipitation >= 0.01 inches for 29-day windows centered on each day of the year (aka DLY-PRCP-PCTALL-GE001HI)
(dly-)snow-50pctl - 50th percentiles of daily nonzero snowfall totals for 29-day windows centered on each day of the year
(dly-)snow-pctall - Probability of snowfall >= 0.1 inches for 29-day windows centered on each day of the year (aka DLY-SNOW-PCTALL-GE001TI)
Valid_day - day of the year for which normal applies (MMDD)
wea_num_type - dly-tmax-normal, dly-tmax-stddev, dly-tmin-normal, dly-tmin-stddev, dly-prcp-50pctl, dly-prcp-pctall, dly-snow-50pctl, or dly-snow-pctall