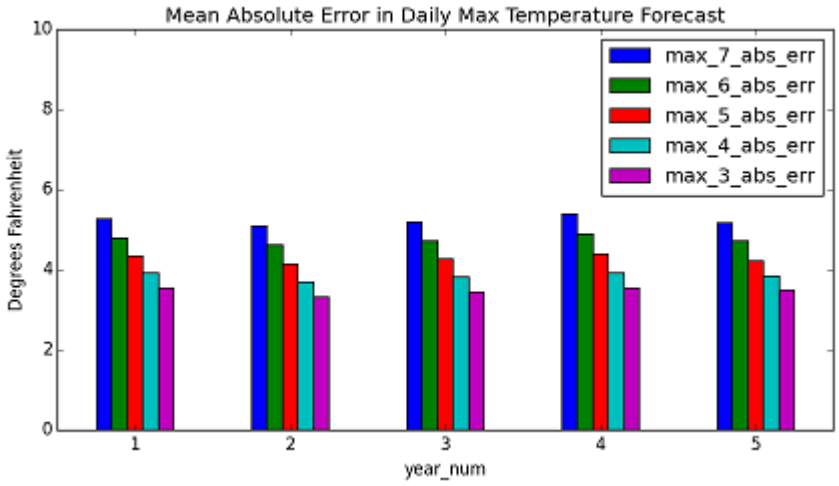
Accuracy of NWS Medium Range Weather Forecasts

5 Years: Jan 8, 2011 – Jan 7, 2016



General Assembly – Data Science class project

Bruce Aker

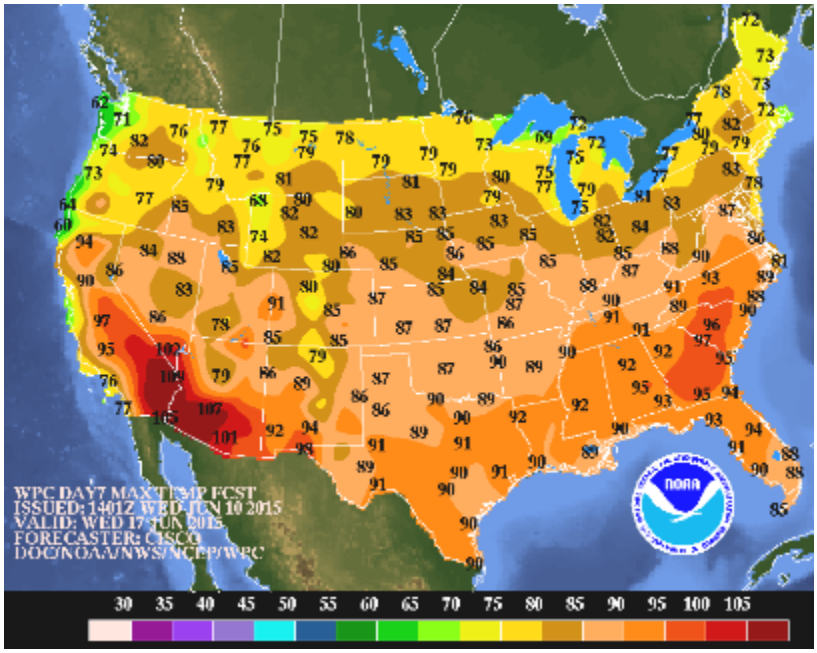
January 21, 2016

Revised February 2, 2016

Question: What is the accuracy of National Weather Service medium range forecasts?

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. Forecasts were then compared with the observations made each day over a five year period.

Maximum temperature forecast issued 6/10/2015 valid for 6/17/2015 (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

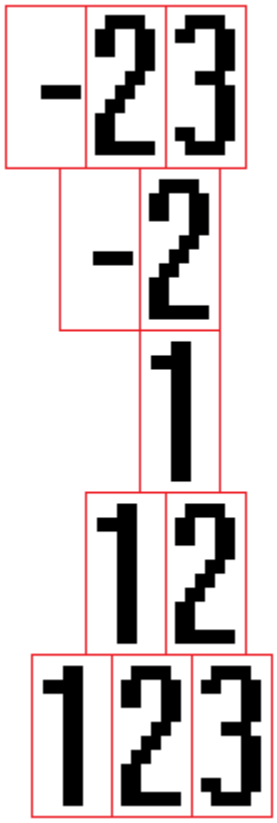
III. Extraction Procedure: the numbers contained up to three characters (within range -99 to 999) from a set of eleven characters (0, 1, 2,…, 9, -). The character font, color (black), size, location was (almost) always the same. There was no other black color, no anti-aliasing and the character overwrote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

Python 9 × 13 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

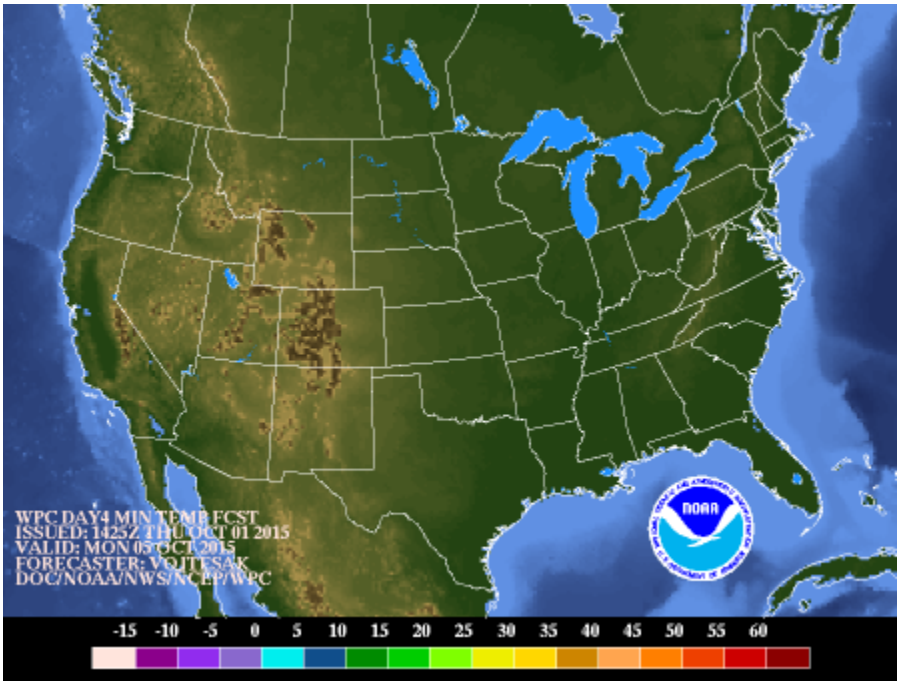
The horizontal position of the number varied depending upon number of characters and whether negative or positive.



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: 2/29/2012, 3/3/2013 (issue date)
- 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 ~September 2012
- Precipitation forecast mostly not available in 2011

V. Obtaining GIF image files:

First step was to download GIF files which are systematically named and publicly available via HTTP. Date encoded as YYYYMMDD.

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:

Second step was to convert to BMP for easier access to pixel array

• 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 12/8/2012:

`DAY7_MIN_20120125_filled.gif  DAY7_MIN_2013012512_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. There were two other similar cases.

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

• Converted from °C and mm to °F and inches

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 were 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:

(Note: valid_date_calcd was calculated from issue_date and forecast_day when forecast numbers were extracted from images).

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
.
.
.
```

med_range_forecast.csv, ~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
.
.
.
```

past_observation.csv, ~626,000 rows, 15 MB

So, using Python and **pandas**, pivot on forecast_type_day and wea_num_type (index = wea_stn_cd, 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
.
.
.
```

forecast_and_obs.csv, 160,688 rows, 18 MB

Exactly the number of rows expected: (365 × 5 + 1) × 88 = 160,688

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

XI. Sample of data:

UNIQUE IDENTIFIER		FORECASTS								OBSERVABLES			
wea_stn_cd	valid_date	max_3	...	max_7	min_3	...	min_7	pop1_3	...	pop1_7	pop2_3	...	pop2_7	prcp	snow	tmax	tmin
KSLC	20130906	90		86	72		70	10		1	13		12	0.000	0.00	98.1	72.0
KBNA	20131001	81		76	63		55	10		12	15		8	0.000	0.00	84.9	64.0
KPHX	20151231	62		59	36		35	0		1	0		1	0.000	0.00	62.1	35.1
KIAH	20120522	89		89	67		69	NaN		NaN	NaN		NaN	0.000	0.00	93.0	72.0
KDSM	20150802	86		82	66		62	35		19	20		15	0.051	0.00	93.0	73.0
KCPR	20150706	71		86	55		59	72		12	46		20	0.039	0.00	70.0	53.1
KWMC	20110815	86		90	50		57	NaN		NaN	NaN		NaN	0.000	0.00	86.0	39.9
KCMH	20140117	31		32	26		29	22		18	27		19	0.091	0.91	35.1	13.1
KACV	20141227	55		55	40		42	0		2	2		11	0.000	0.00	53.1	32.0
KSLC	20130317	50		58	41		43	16		12	17		15	0.000	0.00	52.0	30.0
KDSM	20130828	96		92	76		71	2		15	4		14	0.000	0.00	99.0	73.9
KACY	20150314	53		45	36		32	88		59	91		61	1.130	0.00	55.9	41.0
KACV	20151029	56		60	51		52	32		55	17		38	0.000	0.00	64.0	44.1
KOKC	20130629	97		91	76		71	0		11	2		10	0.000	0.00	93.0	75.9
KCRP	20120930	85		91	70		68	43		19	25		25	0.000	0.00	86.0	72.0
KLBB	20110601	89		91	66		65	NaN		NaN	NaN		NaN	0.000	0.00	95.0	71.1
KTPA	20121120	75		71	61		54	4		9	7		9	0.000	0.00	75.9	59.0
KBOS	20131212	22		27	17		18	5		4	1		7	0.000	0.00	27.1	18.1
KABQ	20151124	60		55	31		30	1		4	1		5	0.000	0.00	63.0	30.2
KGRR	20110506	57		58	46		42	NaN		NaN	NaN		NaN	0.020	0.00	64.9	46.0
KMCI	20130707	90		87	71		69	14		5	9		10	0.169	0.00	91.0	70.0
KCRP	20140503	84		82	60		61	2		4	3		2	0.000	0.00	93.0	53.1
KSHV	20140731	80		82	68		66	45		25	54		35	1.118	0.00	75.9	68.0
KDFW	20120707	98		99	77		77	NaN		NaN	NaN		NaN	0.000	0.00	100.9	73.9
KIAH	20130521	89		89	73		72	12		17	12		24	0.000	0.00	89.1	77.0
KBGR	20130702	76		77	65		66	64		57	63		68	0.429	0.00	66.0	57.0
KLBF	20120130	58		45	24		17	NaN		NaN	NaN		NaN	0.000	0.00	68.0	18.0
KJAX	20140520	83		88	63		63	0		5	1		4	0.000	0.00	79.0	59.0
KLAX	20120817	75		77	66		66	NaN		NaN	NaN		NaN	0.000	0.00	84.0	69.1
KMIA	20151211	80		78	68		70	15		30	11		26	0.000	0.00	82.9	70.0

XII. Missing data:

Some columns of counts of missing data (forecasts, observables) by year_num (and "out of" column), year_num = 1 is 20110108 thru 20120107

	FORECASTS				OBSERVABLES				
year_num	max_3	min_7	pop1_3	pop2_7	prcp	snow	tmax	tmin	out_of
1	21	21	32120	32120	4	6290	16	14	32120
2	123	109	22644	22997	1	2283	89	84	32208
3	88	88	88	88	3	2176	9	9	32120
4	0	0	0	0	9	2514	12	12	32120
5	0	0	1	0	60	2646	61	64	32120

XIII. Analysis:

For daily maximum temperature forecast, calculate error. A random sample:

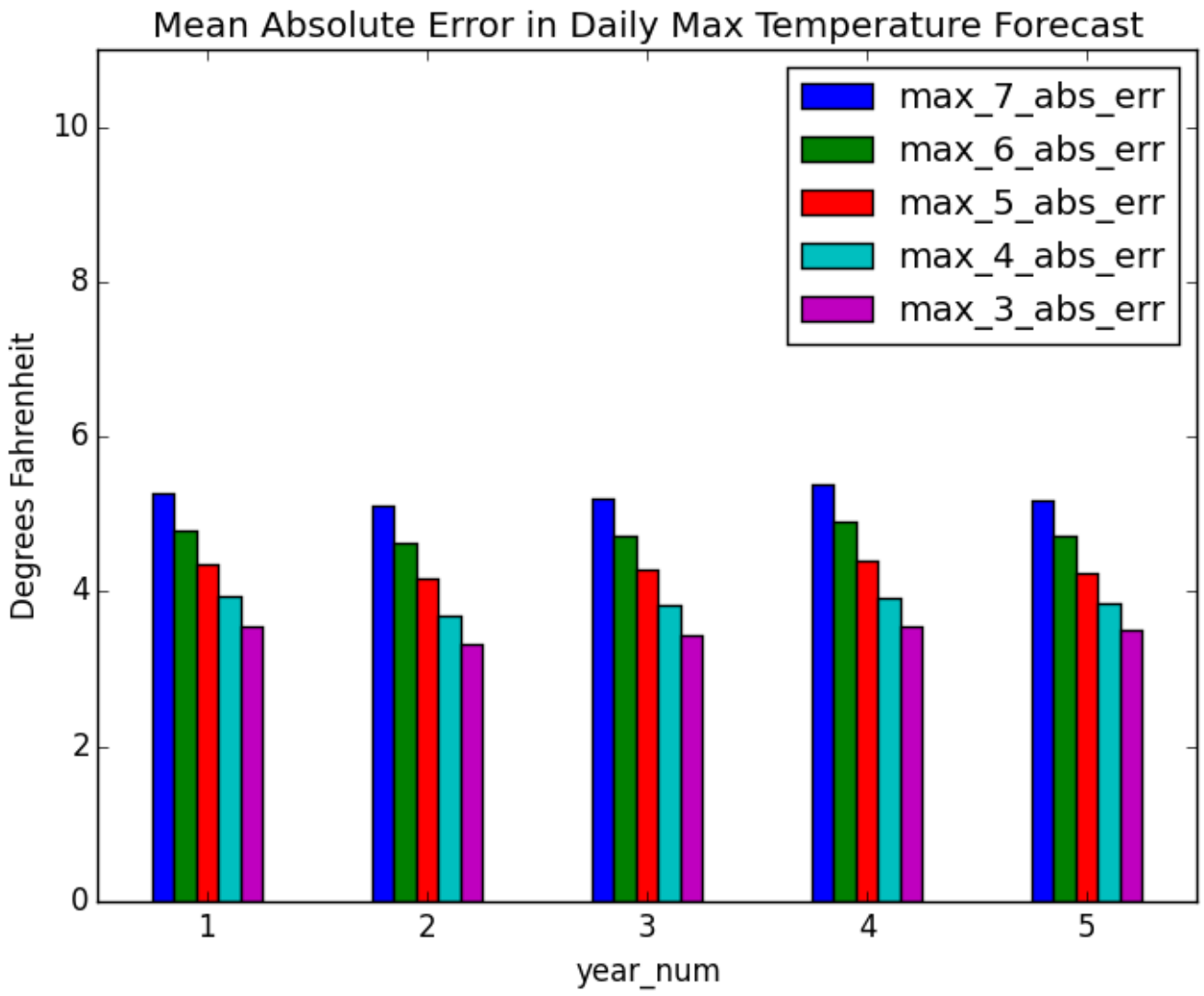
wea_stn_cd	valid_date	max_7_err	max_6_err	max_5_err	max_4_err	max_3_err
KJAX	20130201	-1.0	0.0	-3.0	-2.0	1.0
KGEG	20130321	-2.1	-2.1	0.9	-0.1	-2.1
KFMY	20130323	-6.0	-5.0	-2.0	-4.0	-4.0
KORF	20121021	5.1	2.1	1.1	1.1	1.1
KUIL	20140201	-7.0	-7.0	-7.0	-7.0	-7.0
KDSM	20140814	-4.9	-7.9	-5.9	-5.9	-4.9
KLBF	20121202	-7.1	-7.1	-3.1	-4.1	-4.1
KLAX	20151024	-14.1	-13.1	-12.1	-10.1	-10.1
KLEX	20150210	9.0	5.0	8.0	12.0	9.0
KBNO	20150705	10.0	8.0	3.0	4.0	3.0
KPHX	20150116	-4.9	-4.9	-3.9	-2.9	-1.9
KLAX	20131128	-7.1	-8.1	-9.1	-7.1	-8.1
KDFW	20140829	-1.9	-1.9	0.1	0.1	-3.9
KWMC	20140206	4.0	3.0	-9.0	0.0	-2.0
KIAH	20130412	0.0	-4.0	-5.0	-3.0	-2.0
KOKC	20130528	8.0	5.0	5.0	4.0	3.0
KABE	20150314	0.0	4.0	0.0	5.0	9.0
KALB	20151215	-17.0	-14.0	-11.0	-10.0	-10.0
KEYW	20131001	-5.0	-3.0	-4.0	-3.0	-3.0
KMLI	20150425	7.9	7.9	6.9	5.9	1.9
KRST	20130929	-2.0	-5.0	-5.0	-3.0	-3.0
KEYW	20140609	-2.1	-2.1	-4.1	-5.1	-5.1
KUIL	20141103	-1.0	-1.0	0.0	1.0	0.0
KUIL	20131022	-8.0	-8.0	-8.0	-9.0	-5.0
KDSM	20140421	-5.9	-1.9	-1.9	-0.9	-0.9
KABE	20150930	-2.1	-3.1	-5.1	-7.1	-8.1
KBTV	20121010	4.0	2.0	1.0	5.0	7.0
KROA	20141030	5.0	3.0	0.0	-1.0	-1.0
KMIA	20130217	6.0	9.0	7.0	5.0	2.0
KBIS	20121203	-6.1	-9.1	-11.1	-8.1	-7.1

Positive number for error means forecast was too high

(Hmmm, a lot of negative numbers...)

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

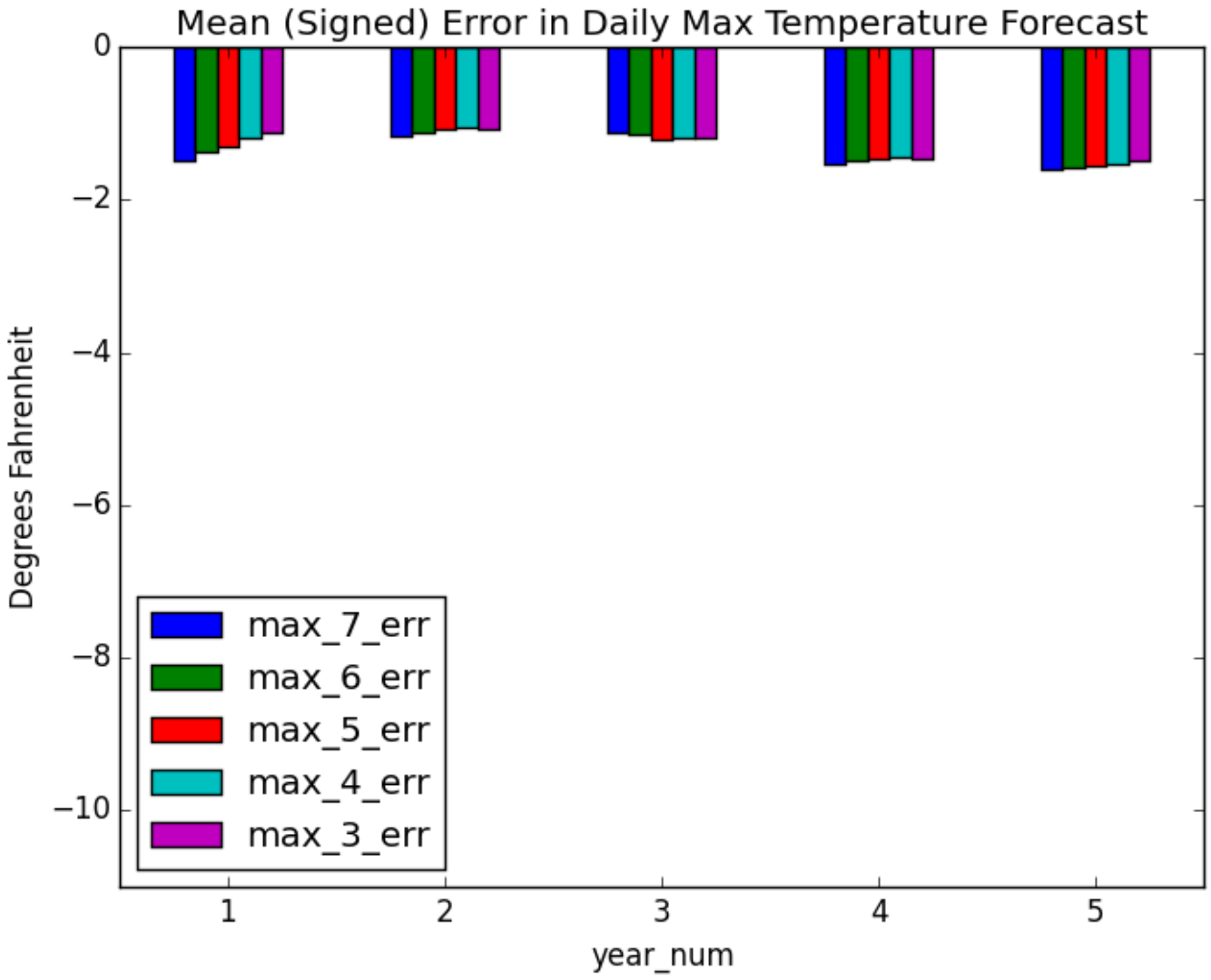
(year_num = 1 is 1/8/2011 – 1/7/2012)



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

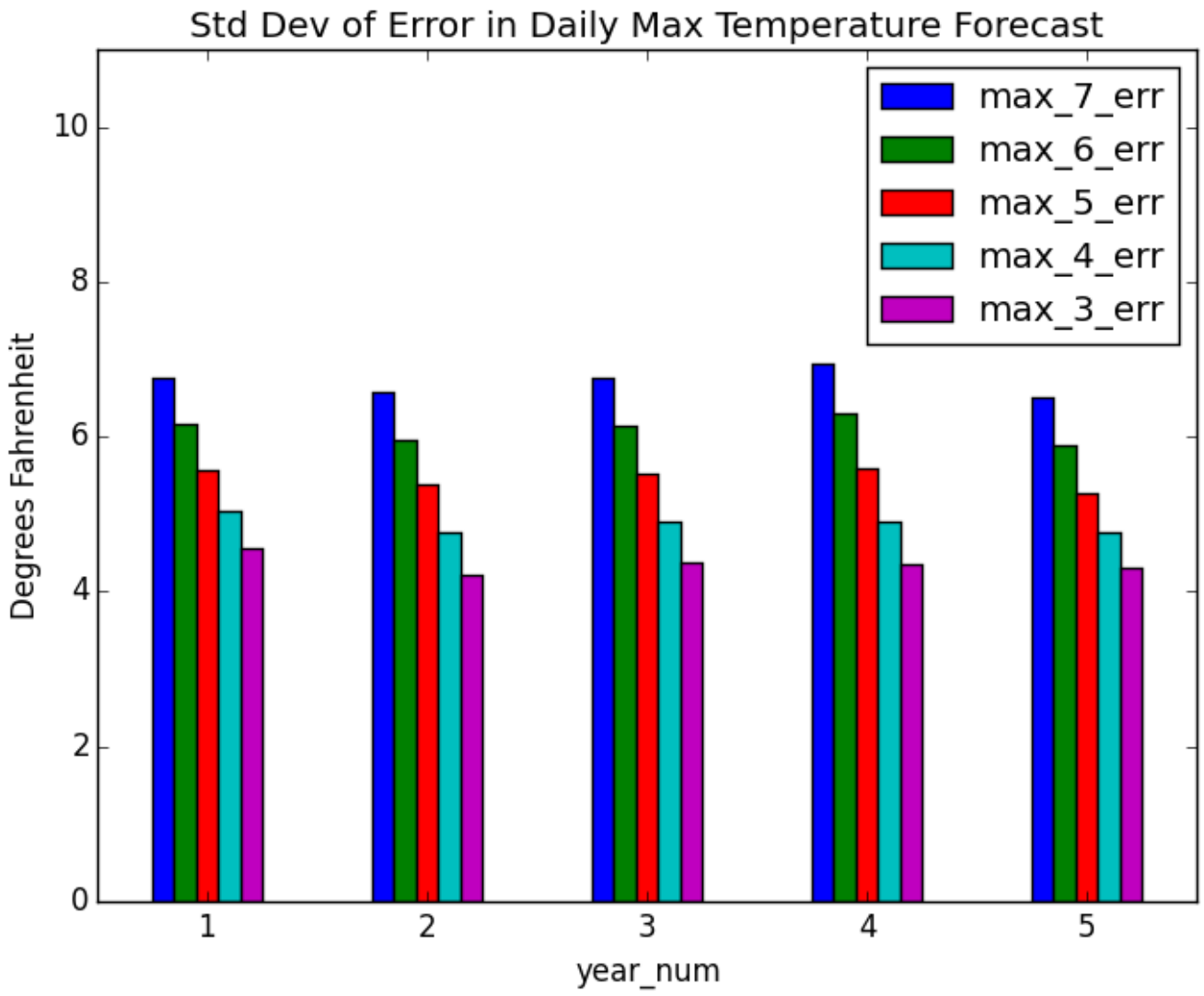
(systematic error, negative number means forecast too low)



Forecast for daily maximum temperature are systematically too low.

Calculate standard deviation of Forecast minus Observed and plot by year_num

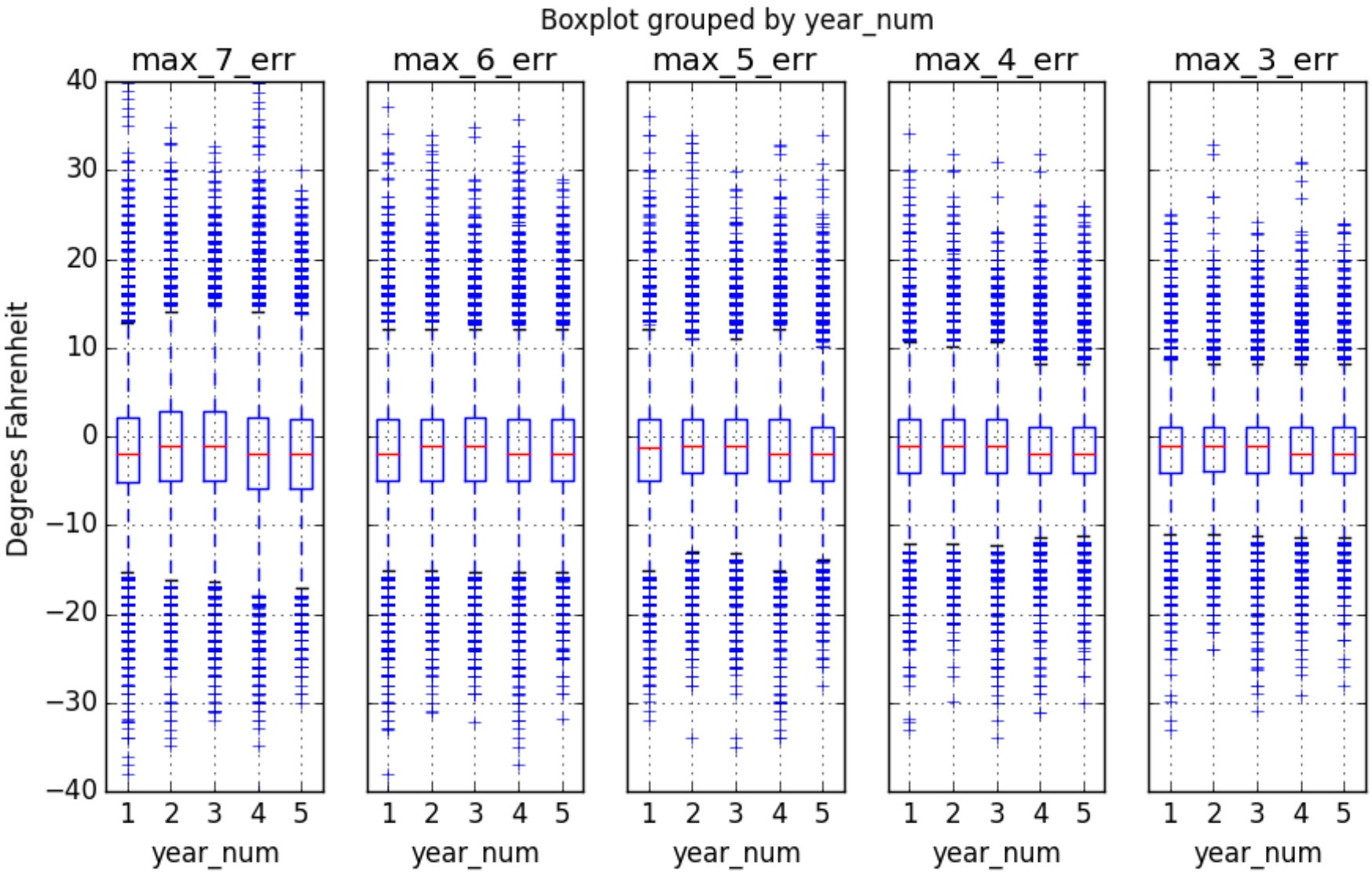
(random error)



3 day forecast is more accurate than the 7 day forecast (obviously).

Box Plot(s) summarizes the data best (shows location and variation)

N.B. medians all below zero

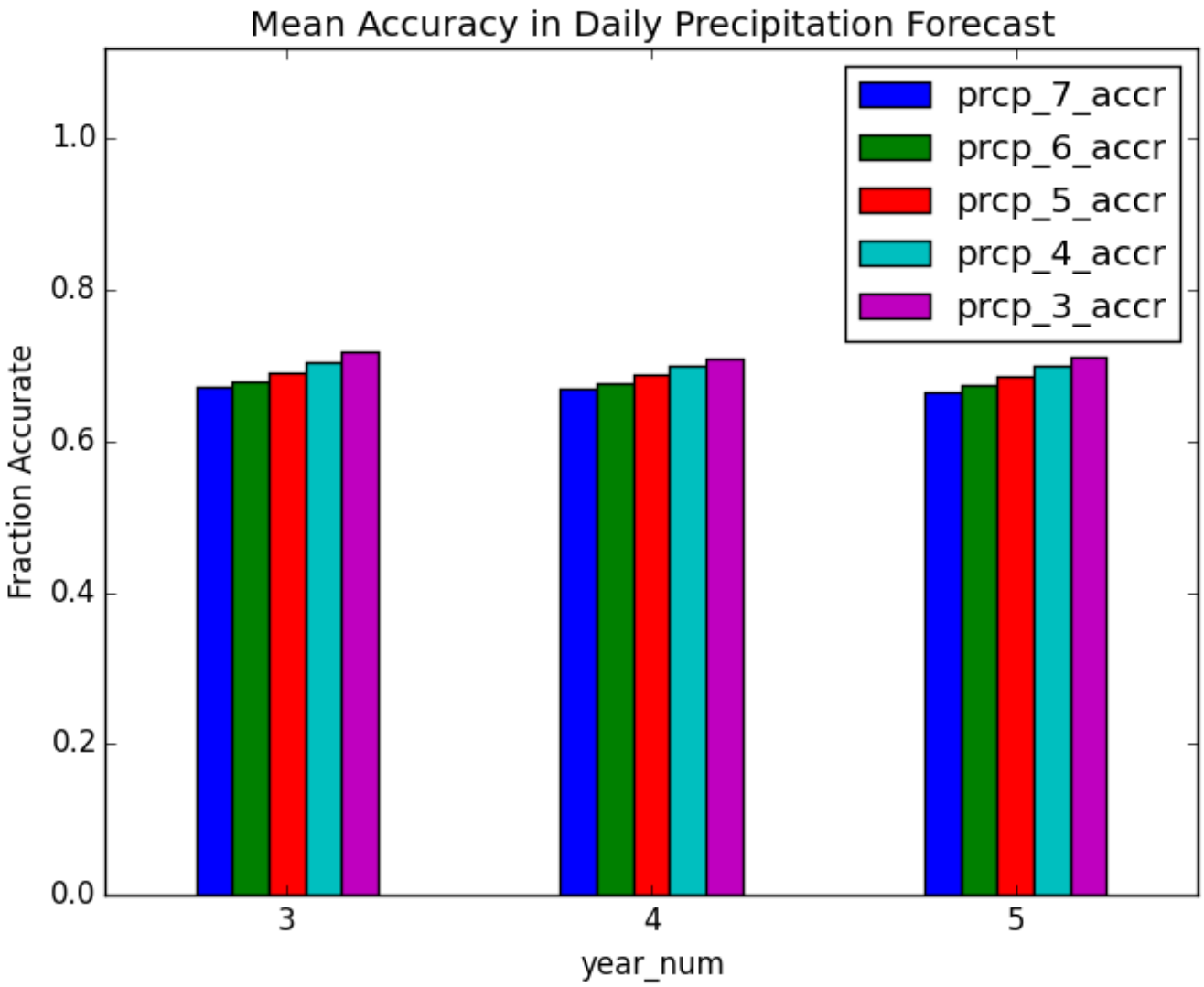


Median error is consistently about -1.0 to -1.5°F

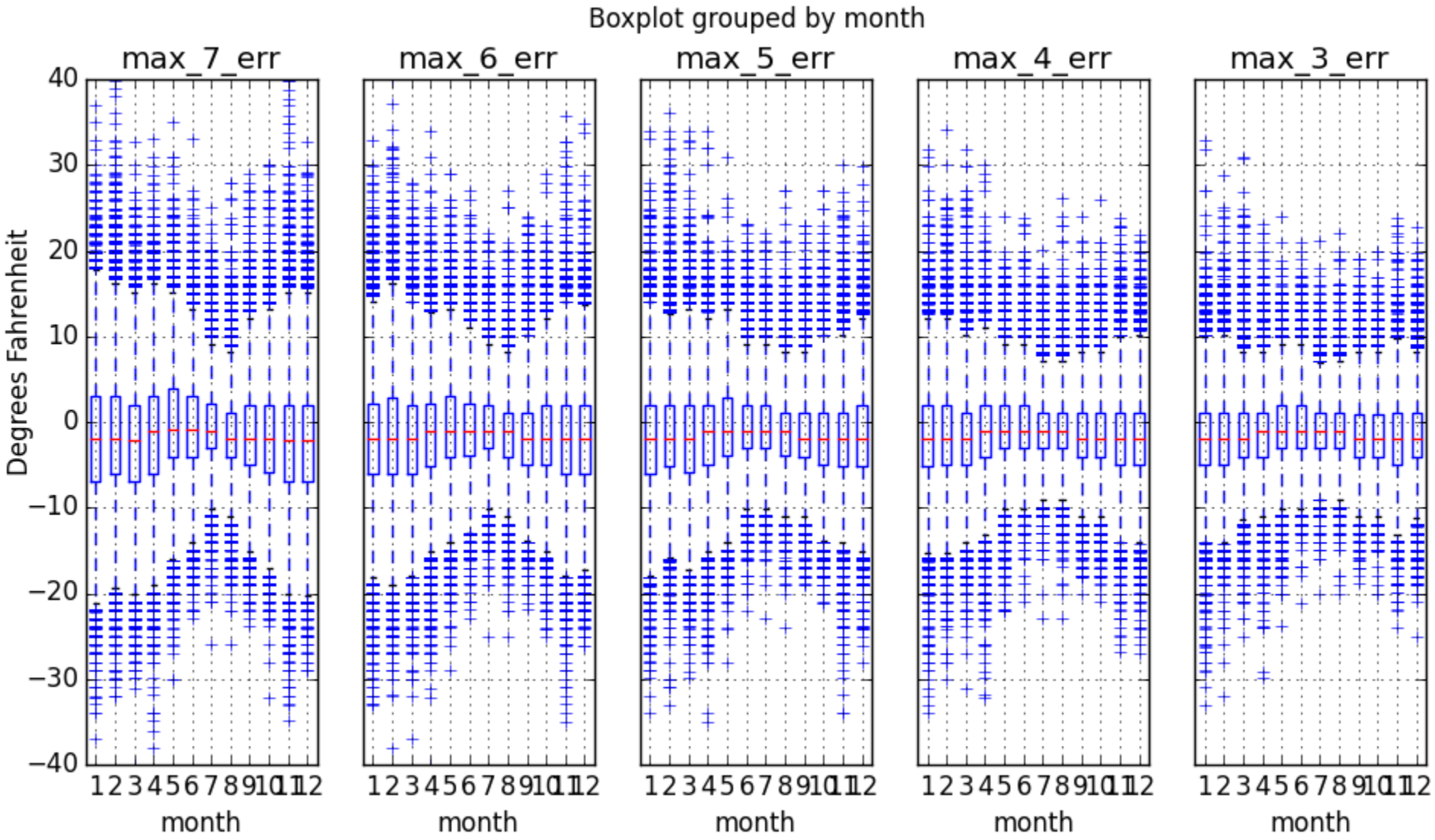
If Probability of Precipitation <= 50% and no rain or snow then score 1,

Or if Probability of Precipitation > 50% and measureable rain or snow then score 1,

Otherwise score 0

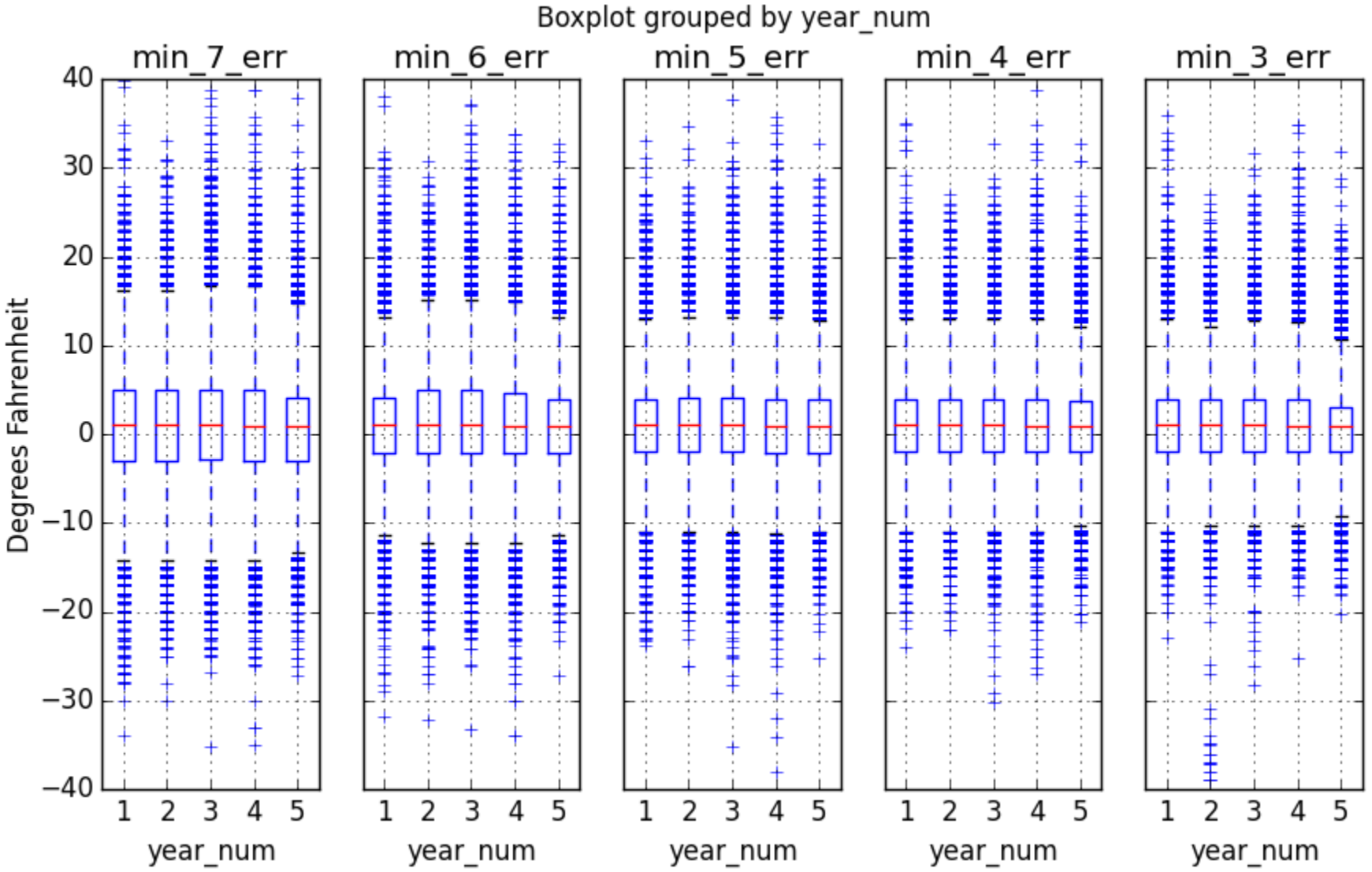


Error in daily maximum temperature forecast by month (all years)



Forecasts in the summer are generally better.

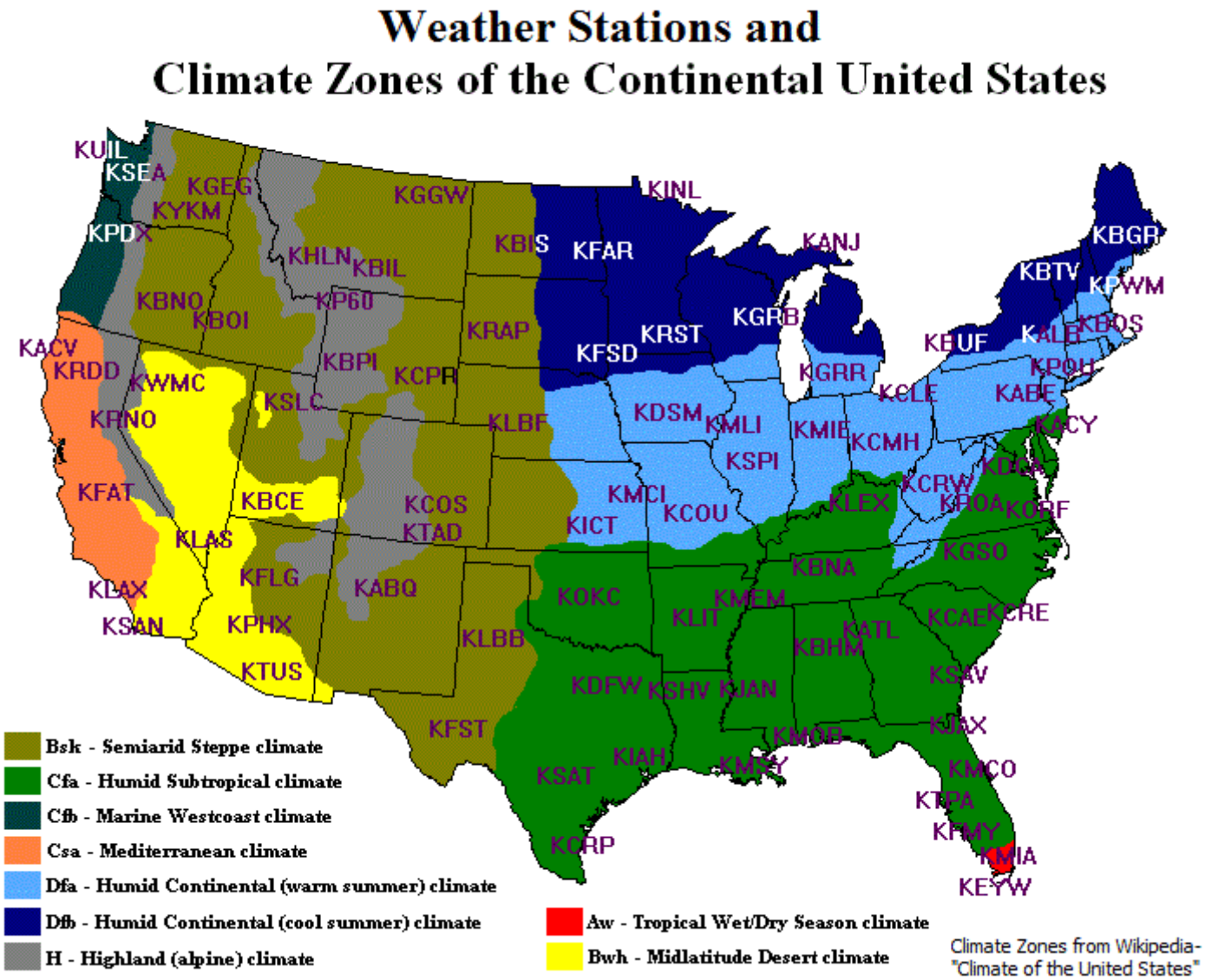
Error in daily minimum temperature forecast by num_year



Medians are all above zero (~1°F)

XIV. To do:

Analyze errors by climate zone



XV. More To Do:

1. Investigate outliers

2. ✓ Investigate missing data – No wea_stn_cd/valid_date combinations are missing, amounts of missing data for each data type is determined

3. Can some missing snow values be reasonably replaced by zero?

4. ✓ Look for duplicate data – all wea_stn_cd/valid_date combinations are unique

5. ✓ Look for non-ASCII characters – all characters ASCII, between ‘,’ and ‘x’, plus CR/LF pairs, no spaces, no tabs, no blank lines

6. ✓ Analyze errors by season, month of the year, etc.

7. 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)

8. Obtain precipitation forecast data from the earlier smaller image sizes, earlier smaller font, maybe using Tesseract OCR

9. 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
year_num – number of year starting with Jan 8, 2011 thru Jan 7, 2012
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
NB – Nota Bene (note well)
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
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