SAS® GLOBAL FORUM 2018

USERS PROGRAM

Reading, Wrangling, Visualizing and Modeling the Surface Temperature of the Great Lakes

Laura Ring Kapitula, PhD
Associate Professor
Department of Statistics
Grand Valley State University

April 8 – 11 | Denver, CO
#SASGF

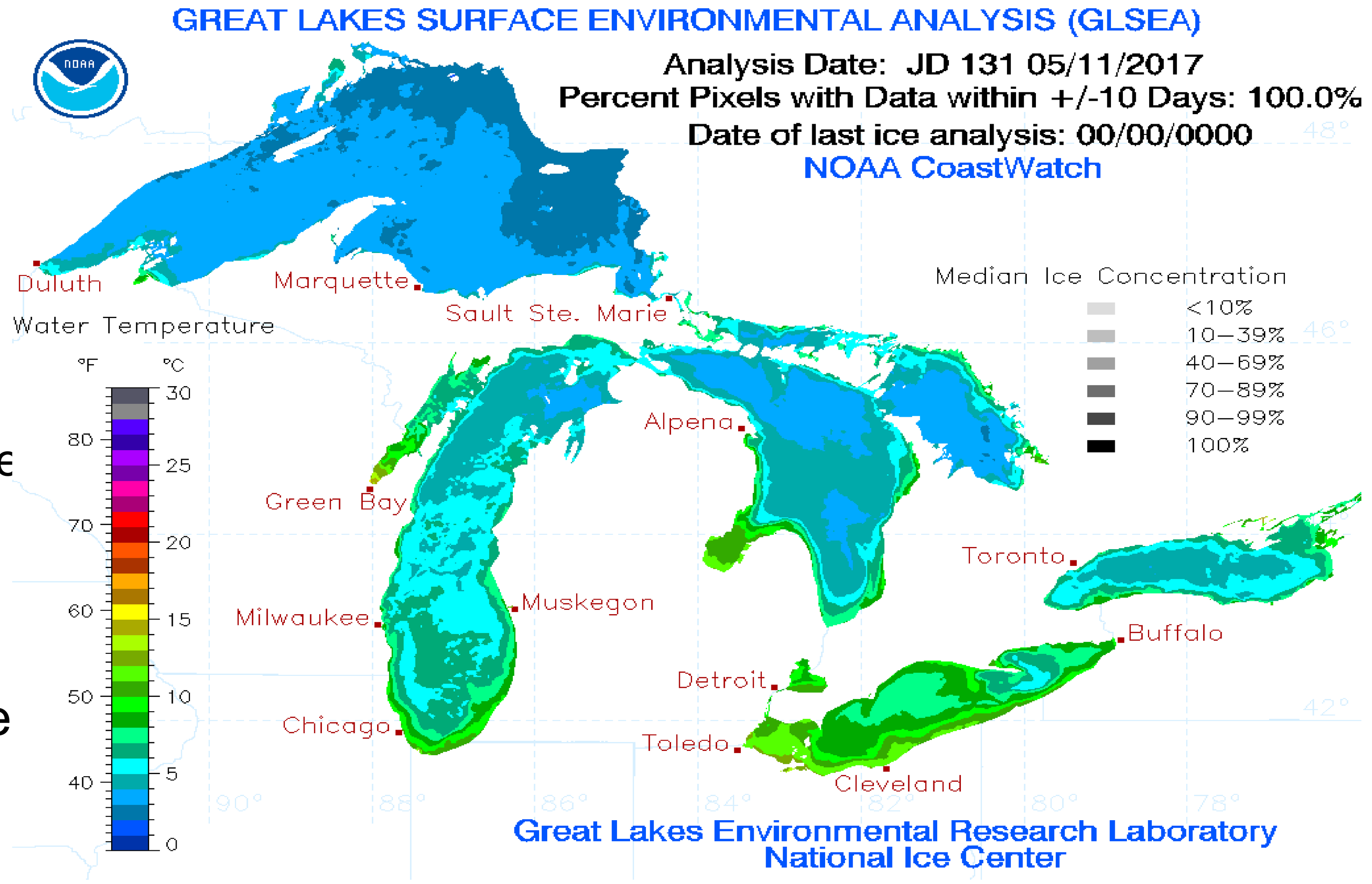
Reading, Wrangling, Visualizing and Modeling the Surface Temperature of the Great Lakes

Laura Ring Kapitula, PhD   Grand Valley State University, Department of Statistics

Grand Valley State University

The Great Lakes

- The Great Lakes are the largest group of freshwater lakes on the planet.
- They contain 21% of the world's fresh water and 95% of the United States surface area fresh water.
- The Great Lakes are diverse.
  - Lake Superior, the largest fresh water lake in the world, is cold and deep.
  - Lake Erie is relatively warm and shallow.
- Surface temperature of The Great Lakes, lake ice amounts and water levels may be important climate change indicators.



"With its rich tradition of agricultural production, commercial and sport fishing, industrial manufacturing, and tourism and recreation, the Great Lakes' economic activity surpasses that of most developed nations. " – Save our Great Lakes

Learning Goals for Student Activities

- Read text data directly from the internet
- Work with Julian Dates and learn how SAS stores dates
- Calculate lags and temperature anomalies.
- Better understand variability, sources of variability and how measurement error impacts variability
- Use SAS by group processing.
- Make attractive reports
- Data visualization
- Data concatenation, merging and restructuring
- Understand map polygons, how to download polygon data for the lakes and use in a statistical package to make a choropleth map and map animation.
- Introduce ideas of predictive modeling and model selection.

The Statistical Computing Class

- About 25-30 Students per section
- Goal is to teach SAS programming and basic ideas of statistical computing and programming.
- Only prerequisite is Introductory Statistics but many students have a broader background.
- Taught in a computer lab.
- Highly interactive.

Reading

- National Oceanic and Atmospheric Administration (NOAA) has a variety of satellite data products. See http://coastwatch.noaa.gov/cw_html/SatelliteDataProducts.html
- CoastWatch is a nationwide NOAA program. The Great Lakes Environmental Research Laboratory (GLERL) functions within this program.
- Using satellite data an average surface temperature is derived for every lake for everyday of the year and are stored in a series of text files, see below:

https://coastwatch.glerl.noaa.gov/ftp/glsea/avgtemps/2016/glsea-temps2016_1024.dat

- Using a filename statement, with a URL statement you can read directly from the data stored on the internet,
- `filename current url "http://coastwatch.glerl.noaa.gov/ftp/glsea/avgtemps/glsea-temps_1024.dat";`

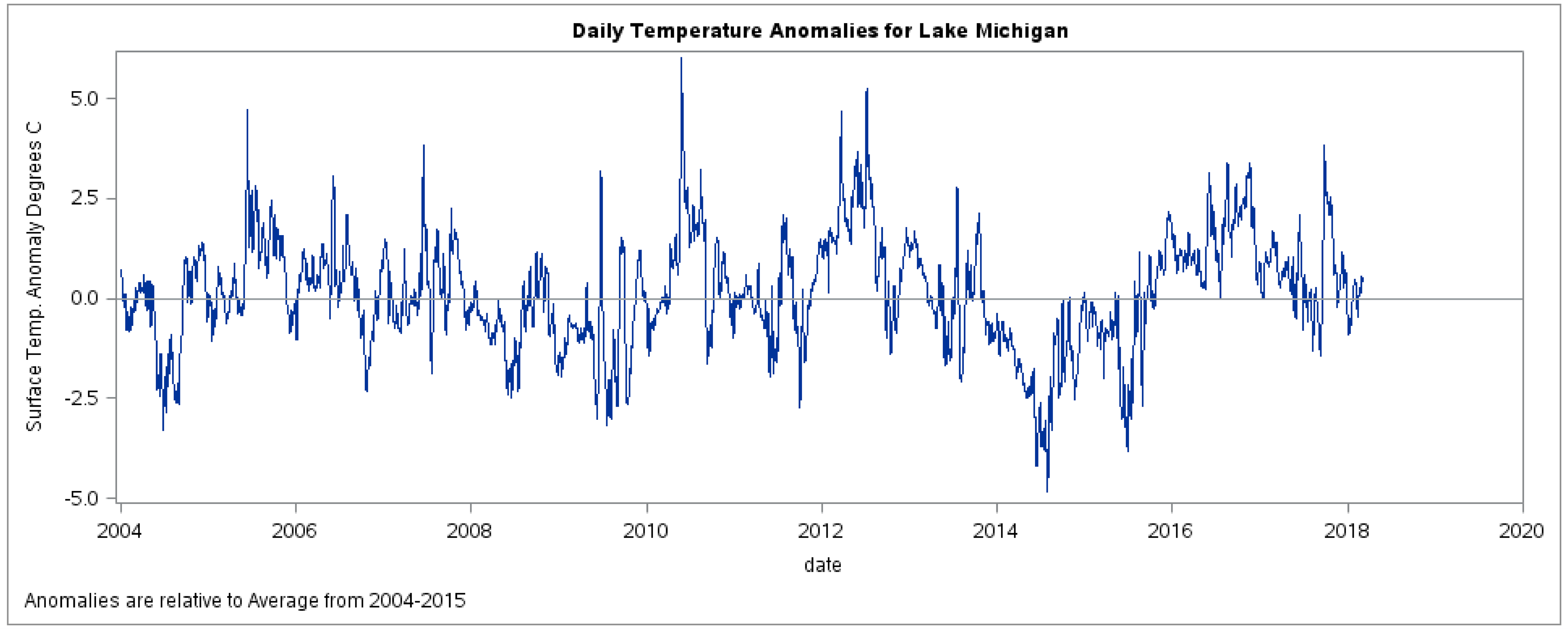
```
 Daily Lake Average Surface Water Temperature
                    From
 Great Lakes Surface Environmental Analysis maps

--------------------------------------------------------
                  Surf. Water Temp. (degrees C)

Year Day    Sup.   Mich.   Huron    Erie    Ont.  St.Clr
--------------------------------------------------------

2016 001    3.94    5.18    5.35    5.37    5.94    3.80
2016 002    3.92    5.15    5.18    5.24    5.77    3.51
2016 003    3.88    5.10    5.03    5.11    5.64    3.34
```

- As students get more skilled they read in all historical data and concatenate to make one big data set.
- Julian dates are converted to SAS date values.

Wrangle and Visualize with the SGplot and SGPanel Procedures

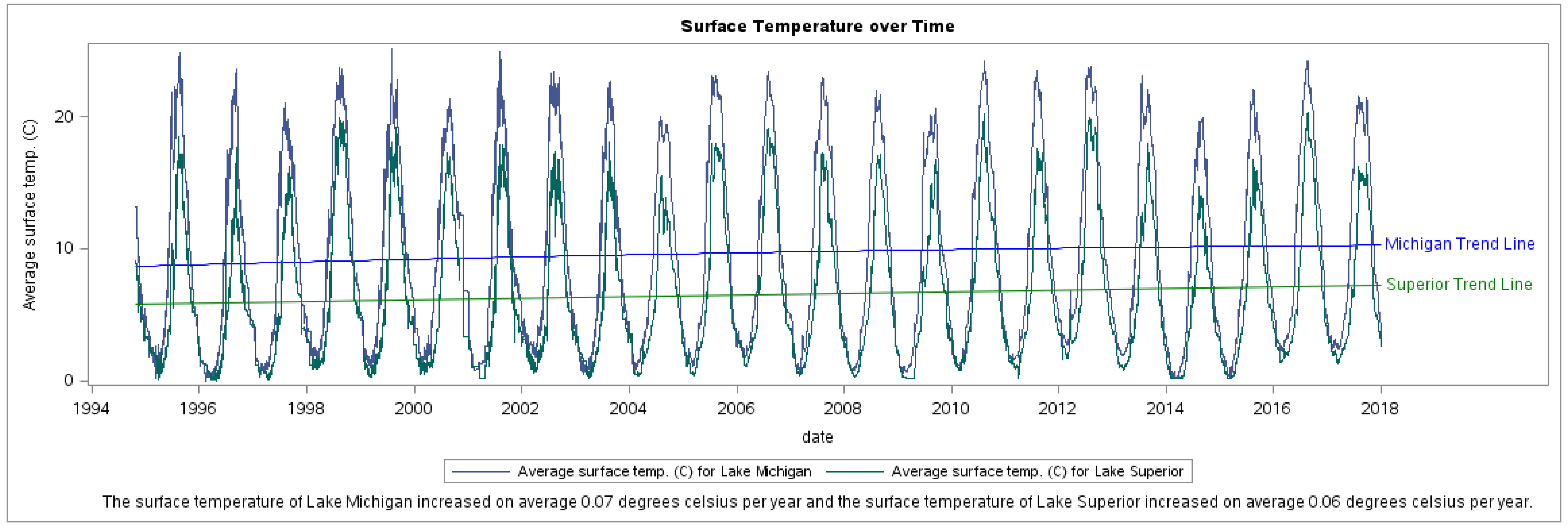
Anomalies

Data are summarized and merged to calculate anomalies and the series statement is used to create create data visualizations:

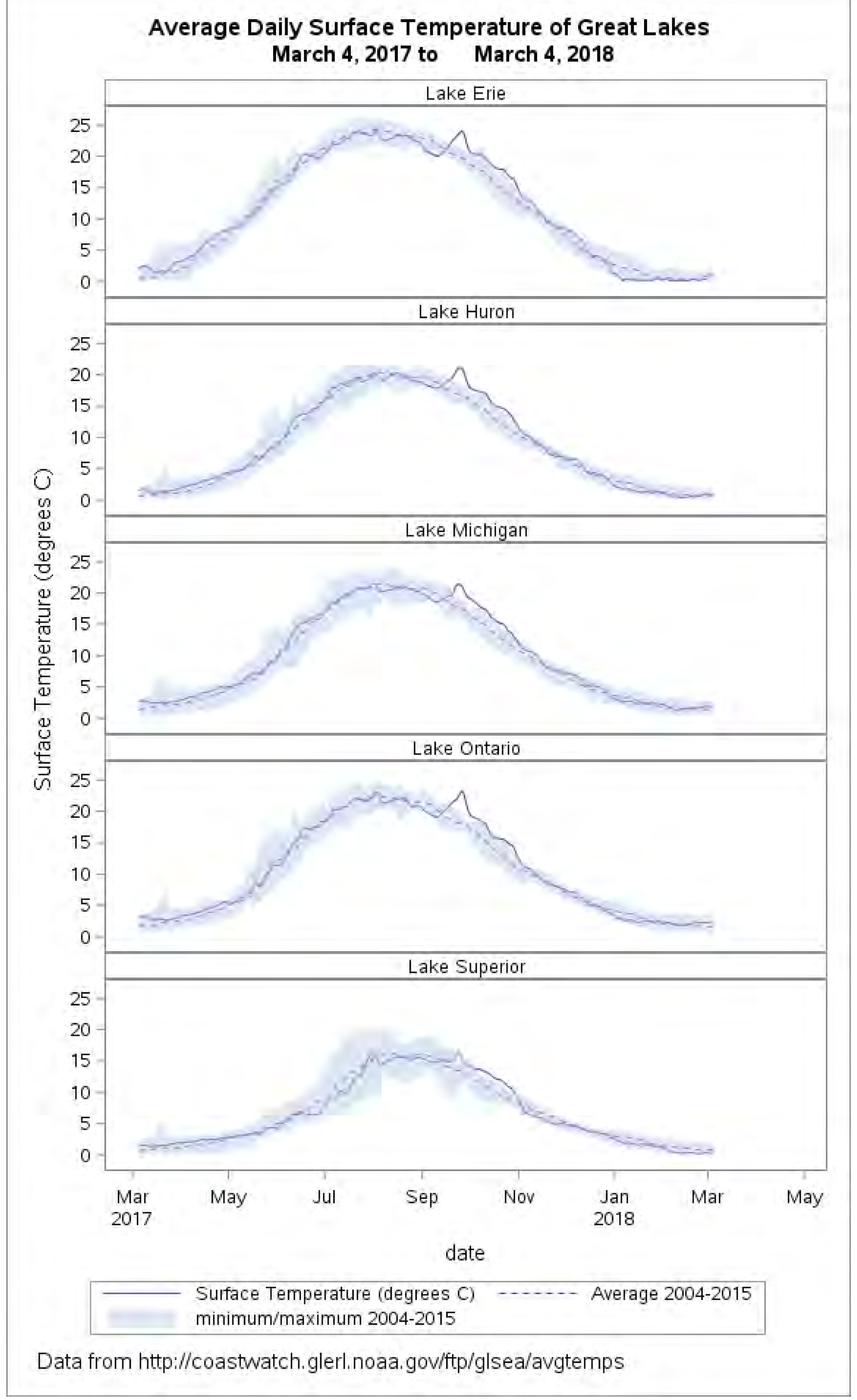


Increasing Average Temperatures

Simple linear regression can be used to estimate how the surface temperature is changing on average over time.

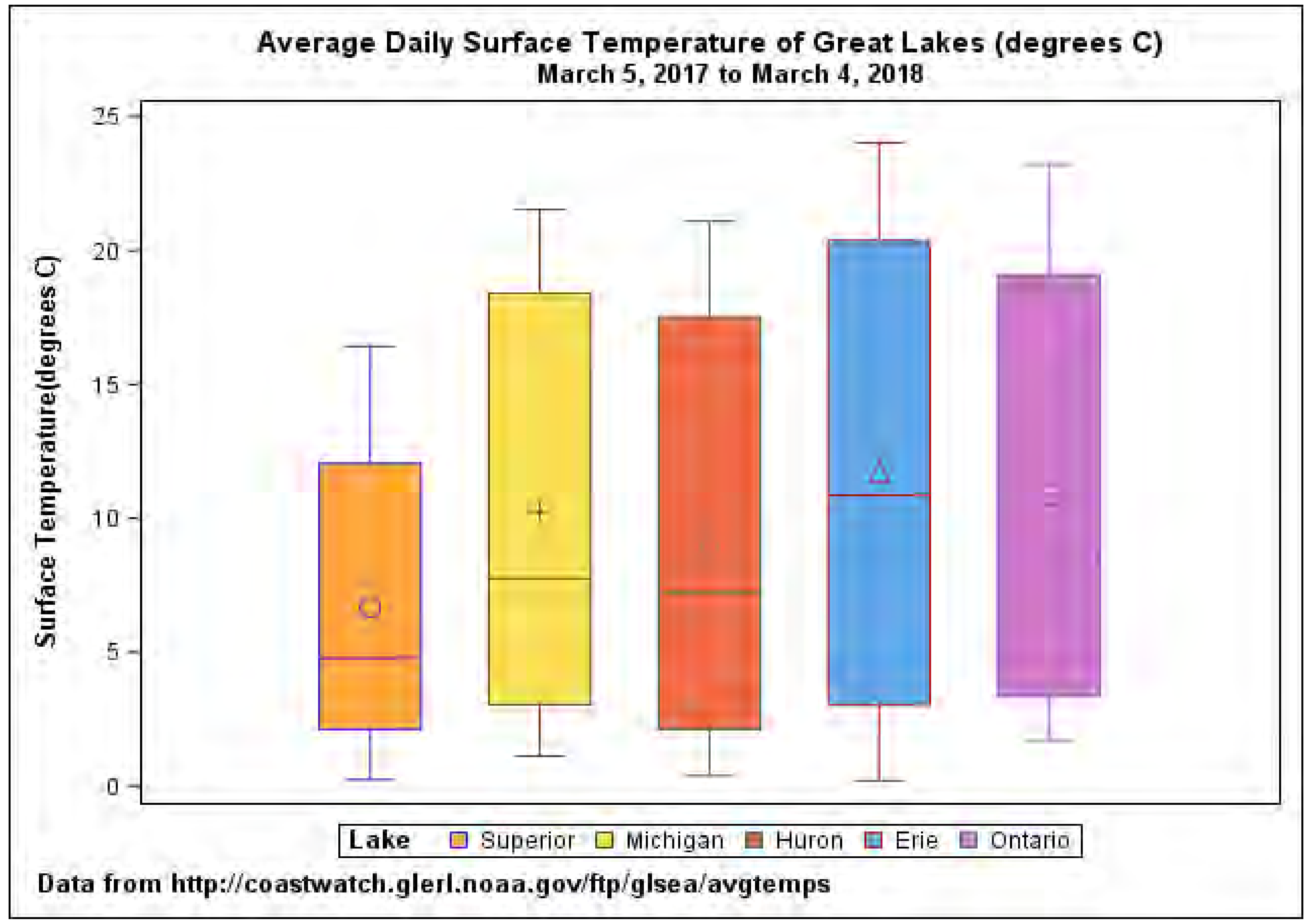


Records Highs in Fall of 2017



Wrangle and Visualize

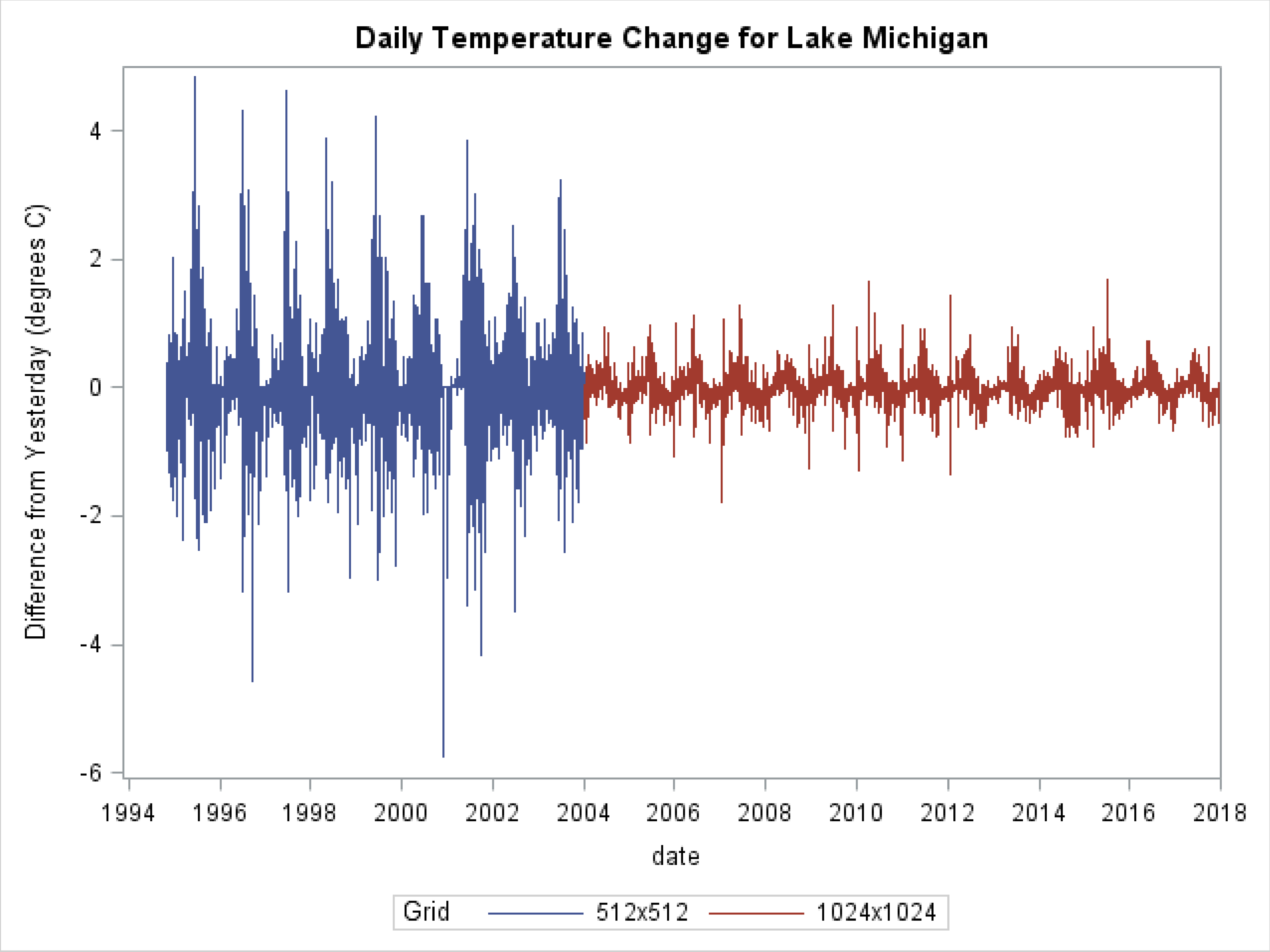
Grouped Box Plots




```
ods text= "5. Transpose the data";
proc transpose data=out.last365
out=skinnylake(rename=(col1=tempc) drop=_name_)
label=lake
 ;
   var sup mich huron erie ont;
   by date;
run;

ods text= "Grouped Box Plots" ;
proc sgplot data=skinnylake;
 vbox tempc / group=lake;
 yaxis label="Surface Temperature(degrees C)";
xaxis label=" ";
label lake="Lake";
run;
```

Lag Plot

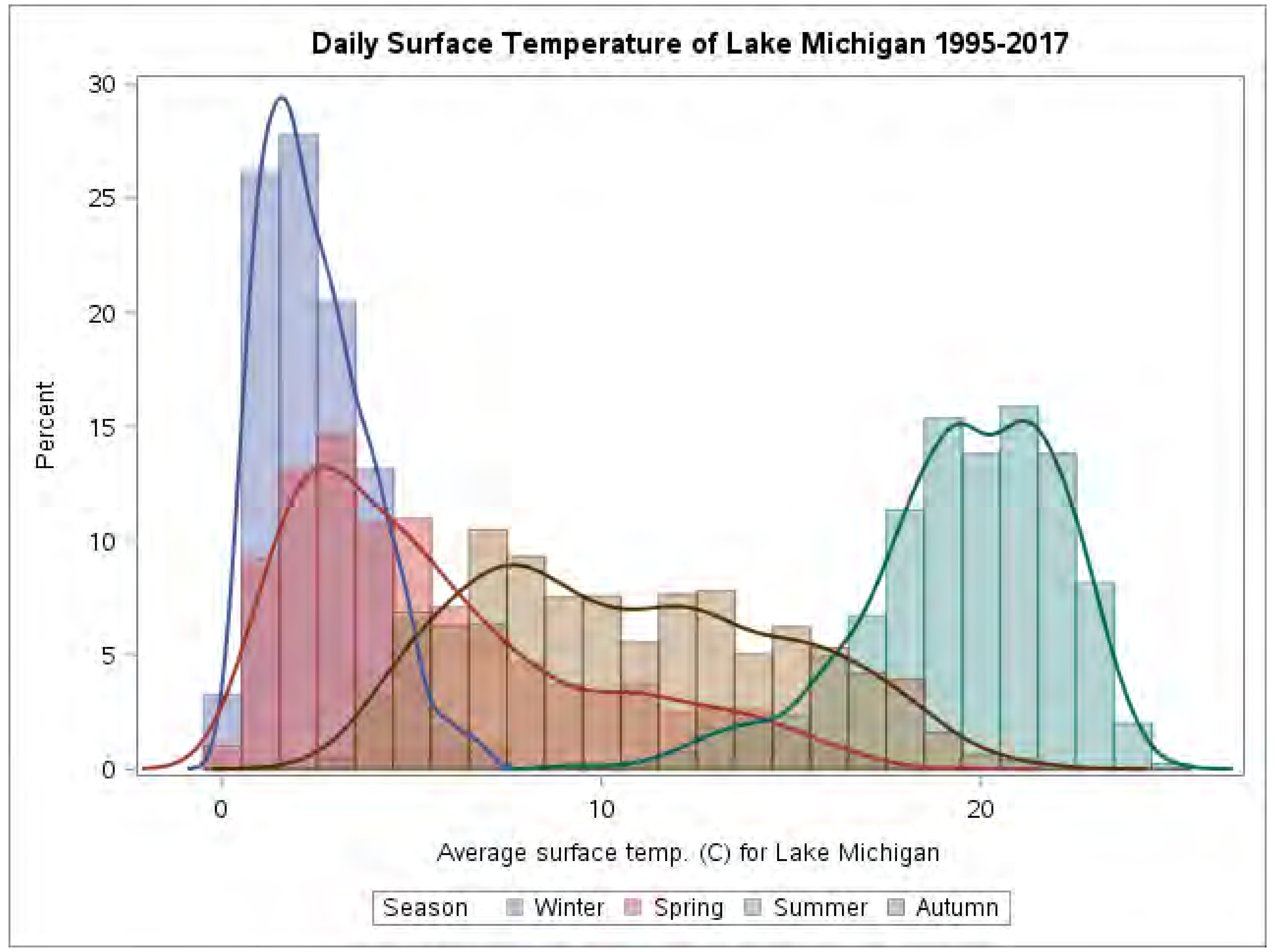
Variability in the average daily surface temperature changes as methodology changes. Higher resolutions means lower variability. What are implications for practice?




```
data gl;
  set glake.greatlakes9417;
  michlagf=lag(mich);
  drop michlagf;
  diffmich=mich-lag(mich);
  if method="1024x1024" and abs(diffmich)>2 then
     name=catx(' ',put(date,mmddyy8.),put(michlagf,4.1),"to",
    put(mich,4.1));
  label diffmich="Difference from Yesterday (degrees C)";
run;
proc sort data=gl;  by date;
ods graphics /reset;
proc sgplot;
  series x=date y=diffmich /group=method
 datalabel=name datalabelattrs=(color=black size=10);
 xaxis offsetmax=.001;
  label method ="Grid";
 title "Daily Temperature Change for Lake Michigan";
  run;  footnote;
```

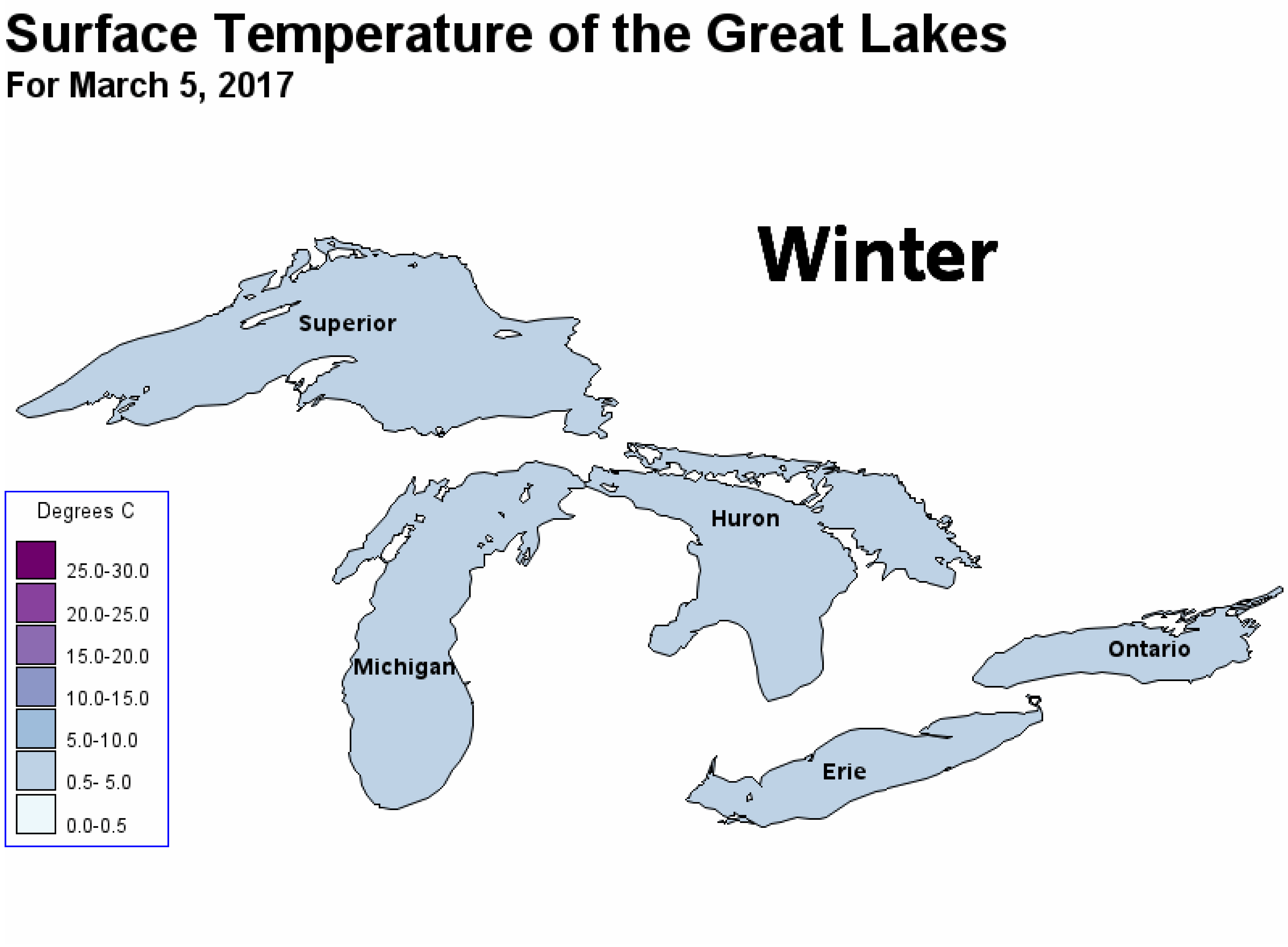
Grouped Histogram

How do the distributions differ? Why is zero the lower bound?
Use formatting to categorize a quantitative variable, use PROC SGPLOT and transparency to create grouped histogram.




```
proc format;
   value season .= "missing"  1-78,355-high= "Winter"
                     79-171= "Spring" 172-264 ="Summer"
                    265-354= "Autumn" ;
                    run;
footnote;
title "Daily Surface Temperature of Lake Michigan 1995-2017";
proc sgplot data=glake.greatlakes9417;
  where year>=1995;
    histogram mich / group=day transparency=0.5;       /* SAS
9.4m2 */
  density mich/ type=kernel group=day; /* overlay density
estimates */
  format day season.;
  label day ="Season";
run;
```

Animated Maps and Polygons



- These maps are created from polygon files not included with SAS GRAPH maps.
- Download polygon files: http://www.naturalearthdata.com.
- Colors selected from Cynthia Brewer's ColorBrewer.org web site.
- To create the maps the data need to be restructured to have one observation for each day/lake.
- BY group processing is used to create the maps and PROC GREPLAY is used to create the animation.

Thank you to Robert Allison at SAS® for his assistance in creating these maps.

Modeling Snowfall and Conclusions

Modeling Snowfall in Grand Rapids, MI

- Use best subset regression to try and predict total seasonal snowfall in a city to the west of Lake Michigan.
- Possible Predictors: land temperature, lake surface temperature and previous years snowfall (going back 5 years).
- For example, about 48% of the variation in season total snowfall was explained using the model below for example:
- $Total\ Snowfall = -22.8 - 3.5\ July\ Temp + 3.8\ August\ Temp - 3.1\ October\ Temp + 5.1\ Mean\ Lake\ Michigan\ Surface\ Temp\ in\ October - 0.5\ Total\ Snowfall\ two\ years\ ago$
- Ideally need a method that takes into account the seasonal nature of the data but linear regression is a good first step.
- Beware of overfitting.
- Many predictors are collinear.
- No clear “best” model.

Conclusions

- Working with data on the Great Lakes provides an interesting and engaging way for students to meet learning objectives in a SAS programing course.
- We see evidence that the Great Lakes are warming.
- The data is simple to understand but complex enough to be realistic.
- In more advanced courses more advanced ideas can be brought into the data analysis.
- It is worth the extra effort to bring real data into the classroom.
- This data can be used in introductory statistics for doing summary statistics or finding correlations and looking at relationships between variables.
- Many other similar data sets are available at NOAA and there are lots of different ways to use the SAS system and real data to better understand our world.

References and Resources

- NOAA Great Lakes Coast Watch https://coastwatch.glerl.noaa.gov/
- Climate Change Indicators: Great Lakes Water Levels and Temperatures https://www.epa.gov/climate-indicators/great-lakes
- Global warming and the Great Lakes. https://www.nwf.org/Wildlife/Threats-to-Wildlife/Global-Warming/Effects-on-Wildlife-and-Habitat/Great-Lakes.aspx
- The Great Lakes. http://www.sustainourgreatlakes.org/about/our-lakes/
- Teaching Great Lakes Science http://www.miseagrant.umich.edu/lessons/lessons/all-data-sets/

SAS® GLOBAL FORUM 2018

April 8 – 11 | Denver, CO
Colorado Convention Center

#SASGF