Abs. ED43C-1118

The Zoo of Metrics: Teaching a Robust Set of Data-Model Comparison Techniques for Space Physics

AGU 100 FALL MEETING
December 9-13, 2019 San Francisco, CA

Michael W. Liemohn and Abigail R. Azari

Department of Climate and Space Sciences and Engineering,
University of Michigan, Ann Arbor, MI USA liemohn@umich.edu

CLASP CLIMATE & SPACE Sciences & Engineering

1. Space physics has a problem

- **In a recent paper, I made this chart**
  - Please notice the variety in the last column

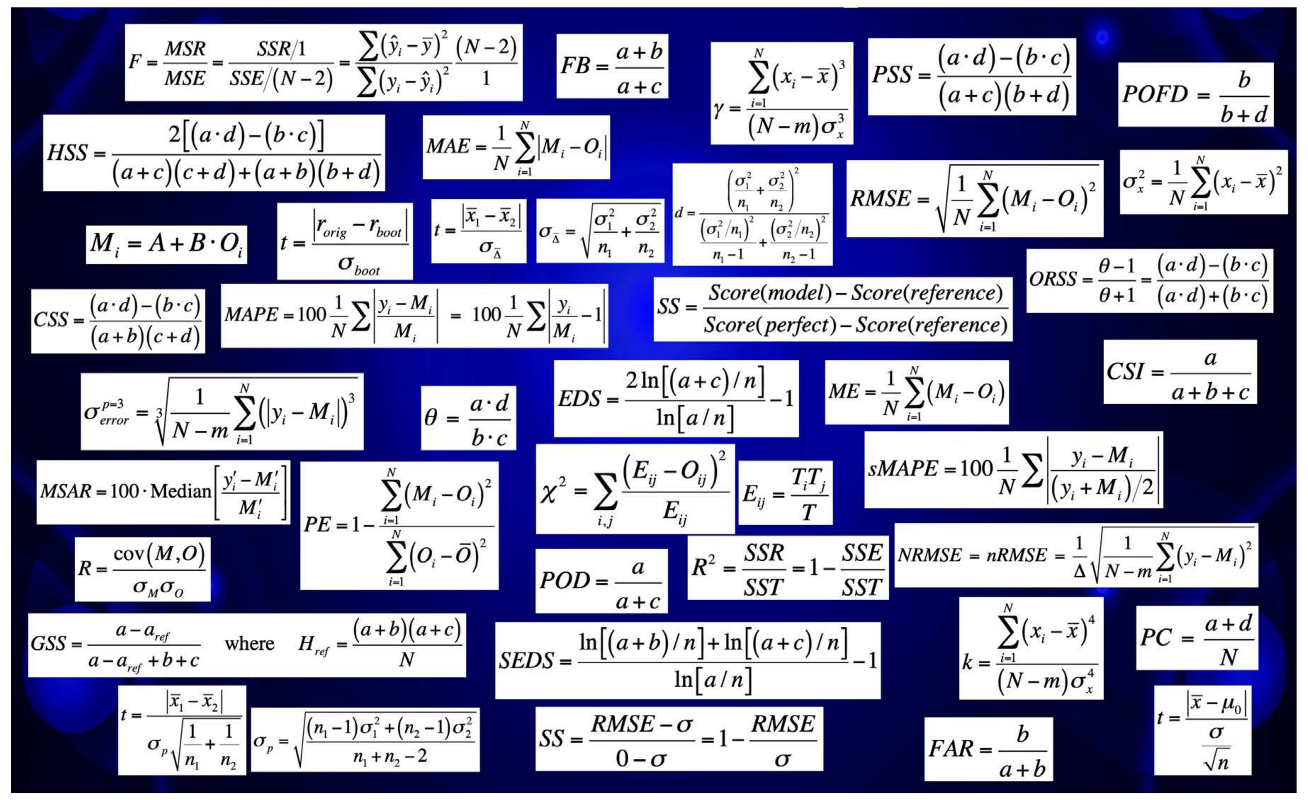
References	Description	Metrics
Boberg et al. (2000)	Time delay neural network	RMSE=0.98, R=0.77
Takahashi et al (2001)	Kp estimation from one or several individual station values	Single station: R between 0.85 and 0.9; 9 stations: R=0.94
Wing et al (2005)	Feedforward backpropagation and recurrent neural network prediction schemes	R=0.94, Gilbert SS=0.2-0.5 for Kp 2 through 6, depending on year
Bala et al (2009) and Bala and Reiff (2012, 2014)	Feedforward backpropagation neural network scheme	3-h lead-time: R=0.77, RMSE=0.8, HSS for KP>6=0.964
Devos et al. (2014)	Prediction of local K-index from Chambon-la-Forêt	R=0.53, ME~0, MAE=0.3, HSS=0.52
Ayala Solares et al. (2016)	Kp with NARX, with both a "sliding window" and a "direct approach" for the input values	3-h ahead: RMSE=0.76, R=0.87, PE=0.76; 24-h ahead: RMSE=0.87, R=0.83, PE=0.68
Wintoft et al. (2017)	Ensemble of time delay neural networks	RMSE=055, R=0.92 (function of year and Kp)
Savani et al. (2017)	Kp prediction from predicted solar wind based on a coupling function empirical formula	POD=0.67, FAR=0, TS=0.6, TSS=0.6
Haiducek et al. (2017)	Kp prediction from SWMF for all of Jan 2005	RMSE=1.1, ME=0.7

Liemohn et al., SWE, Dec 2018

- **This column has many different metrics listed**
  - Our community uses a wide range of data-model comparison formulas
- **I now teach a course at the University of Michigan on data analysis and visualization**
  - **Metrics are a large fraction of the class content!**

2. The zoo of metrics – but we love RMSE!

- **There are many data-model comparison metrics to choose from…**



- **Space physics papers often only use two:**
  - Correlation coefficient (R) and root mean square error (RMSE)
  - Occasionally we use another, like prediction efficiency
- This is barely scratching the surface of what we can explore and hopefully learn from a comparison of observations and models

3. Categories of metrics

- There are several major categories of metrics, each focused on a certain aspect of the fit. Here are a few of **the major categories**:
- **Accuracy:** *How close is the model to the data?*
- **Bias:** *What is the discrepancy between the model and the data?*
- **Precision:** How similar is the clustering tendency in the model and data?
- **Association:** *How well to the model and data values move together?*
- **Extremes:** *How well can the model get the outliers in the data?*
- And the **subsetting categories**, using the above metrics on a portion of either the data or model values:
- **Discrimination:** *How good is the model for a specific range of the data?*
- **Reliability:** *How close is the data to the model values for a specific range of the model output?*
- And a final category, **comparing the metric value to a reference model**:
- **Skill:** *How good is the model at reproducing the data relative to a previous model?*
- **Another dichotomy** is that there are two basic groupings of metrics:
- **Fit performance metrics:** tests the model against the entire data set, usually with a differencing between the model and data values
- **Event detection metrics:** defining events as values beyond some threshold and determining how well the model identifies observed events, without regard to data-model difference

4. Why use more than two?

- Every formula in section 2 can be organized into one of the categories and groupings
- Each metric reveals something different about the model's fit to the data
  - **RMSE** is an **accuracy** measure, but values could be systematically above or below data
  - **R** is an **association** measure of linearity, but values could be very far from the data
  - **ME** (mean error) is a measure of bias but doesn't reveal information about trend, clustering, or extremes
  - I could go on…
- Even from the three just mentioned, **combining them reveals new info beyond any one alone**:
  - R bad, others good: model values jump above and below the data values
  - RMSE bad, others good: model values are close to mean value of data
  - ME bad, others good: model values have the right trend but are offset high/low from data

5. Why lecture you about metrics?

- We have a class at U-M designed for students to explore and learn about data-model comparisons
  - **CLIMATE/SPACE 423: Data Analysis and Visualization for Geoscientists**
- **It's a "zero-to-hero" approach to applied statistics:**
  - Students first learn about processing a single data set (histograms, mean, …), then two data sets (x-y pairs, …)
  - Students learn about simple models based on the data (linear regression, polynomial regression, …) and simple metrics (correlations, chi-squared, ANOVA tables, …)
  - Students then learn about the full suite of metrics described above and the strengths and limitations of each
- **It's a zero-to-hero approach to Python usage as well:**
  - Students are introduced to Jupyter notebooks, using stats packages, opening data sets, and making basic plots
  - Students systematically explore Python commands for all of the stats taught in the class sessions
  - All examples use geophysical data, from the Earth's interior, oceanography, the atmosphere, the magnetosphere, planets, and the Sun
- **Work gets progressively more sophisticated**
  - Homework sets start out very prescriptive, following a set procedure and even being given a template notebook
  - They build into more open-ended mini-projects, using given data sets, that meet certain learning goals
  - Eventually transition to full-scale projects, including written reports and oral presentations, with choice of data

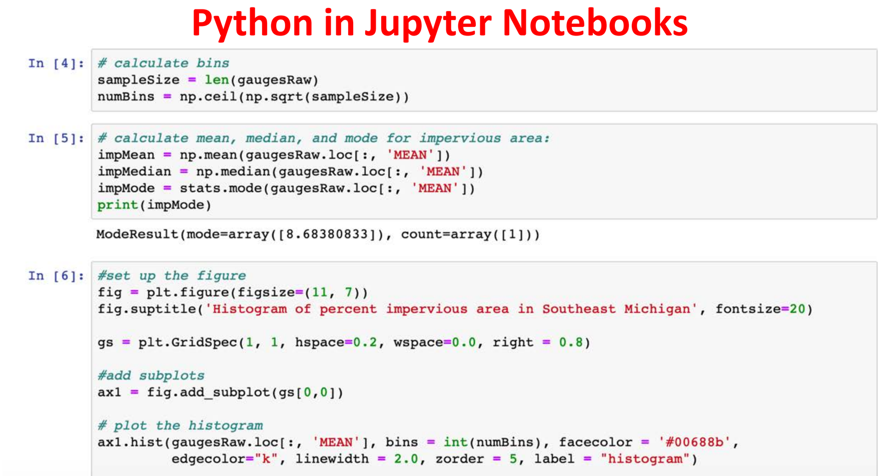
6. Uncertainties!

- **Ascribing uncertainty – perhaps the biggest lesson students learn**
  - It is vital to appreciate the relationship of uncertainty to a value
  - **Critical point: comparing two numbers is meaningless without uncertainties**
- Start this lesson on the first day:
  - Deciding how well we "know" a value
  - Measuring something in the classroom with an unusual unit and ascribe an uncertainty to their length estimate
- Build up to quantitative calculations:
  - Section on uncertainty propagation
  - Content on calculating data set variance
  - Equations for fit coefficient uncertainties
  - Discuss uncertainties on data-model comparison metrics formulas
  - Two half-days on the bootstrap method
- Data-model comparisons: what is "good?"
  - Each metric can usually be compared with a data-set-based value
  - For example, RMSE against standard deviation, "good" when RMSE < σ
  - Discussed and explored for all metrics
- **Students learn to appreciate uncertainty**
  - Extensively worked with it throughout the term

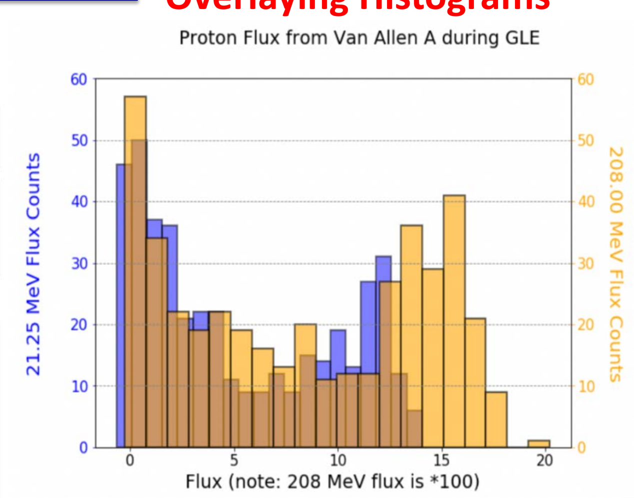
7. Student work from the class

- **A sampling of their final projects**

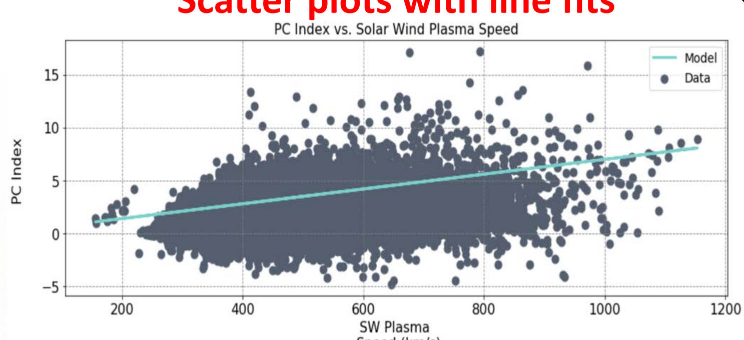
Python in Jupyter Notebooks



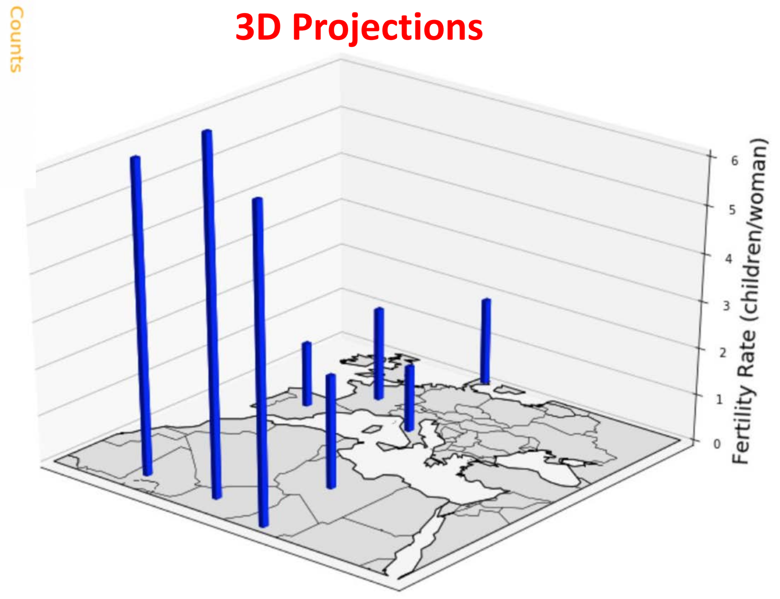
Overlaying Histograms



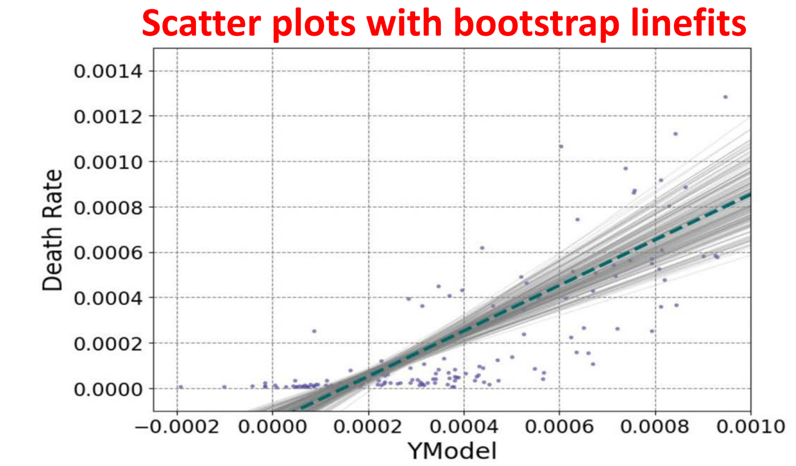
Scatter plots with line fits



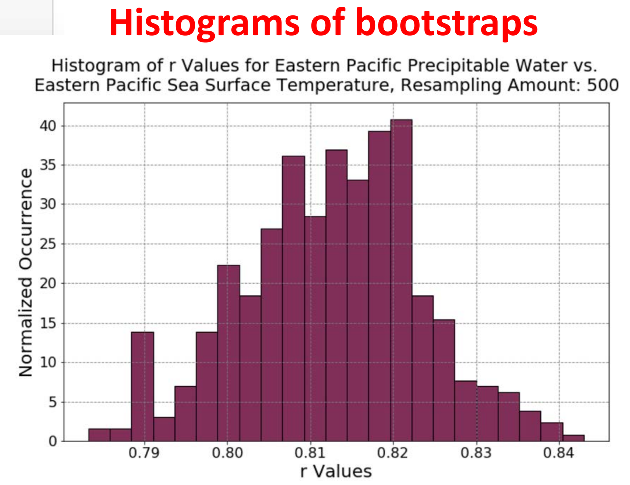
3D Projections



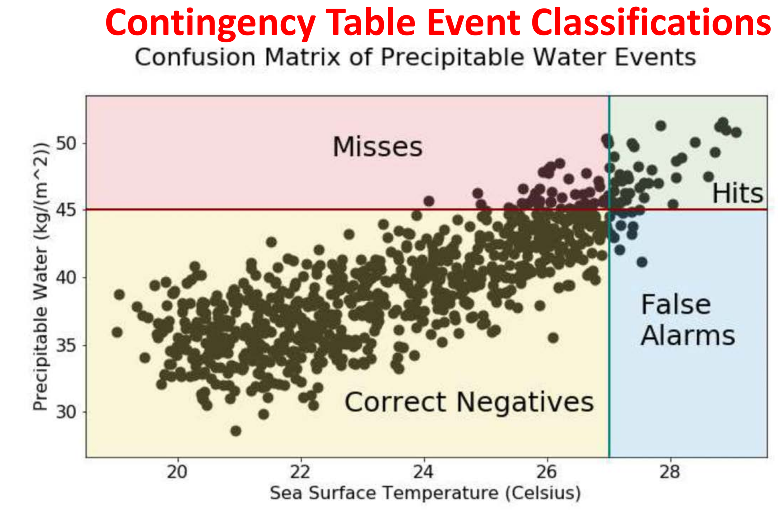
Scatter plots with bootstrap linefits



Histograms of bootstraps



Contingency Table Event Classifications



ROC curves

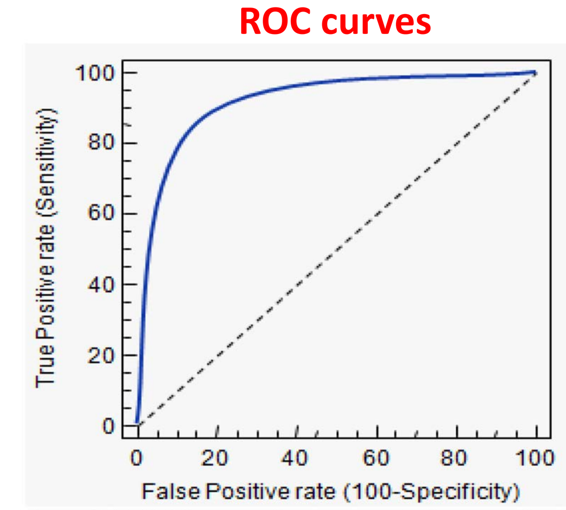


Figure (14): ROC curve of a representative model.

They even spent a day on GIS

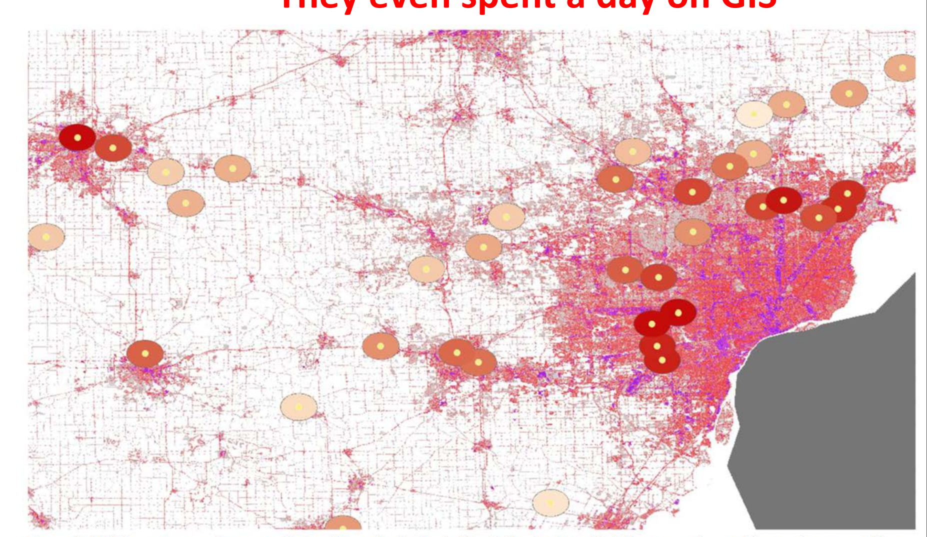


Figure 1: USGS gauges are shown as circles. The color in the buffer indicates how high the percentage of impervious area. The background is the impervious area raster.

8. Summary

- **This is a fun class to teach**
  - Students are engaged and enthusiastic
  - Feeding them the superfood of science and engineering
  - Skill sets learned are applicable to other fields
  - In fact, many students are from other departments
- **Students learn Python**
  - Accessible and open source
  - Jupyter notebooks ease instruction and assignments
- **More on Jupyter in this class:**
  - **See Abby Azari's oral presentation tomorrow "Jupiter with Jupyter"**
  - ED52-06 (11:35 am) in Moscone South Room 216
  - Abby's github site for this: https://github.com/astro-abby/data_vis_statistics_geosciences
- **It's all about uncertainty**
  - Key concept for comparisons
  - **Go to the zoo! (of metrics)**