Physical Sciences: Functional Data Analysis for Rain Rate Statistics

Paul M. Aoki <aoki@acm.org>, Team #721

Problem / motivation

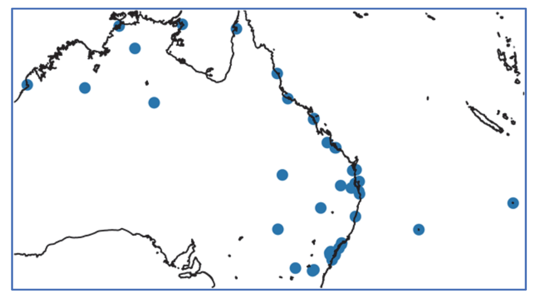
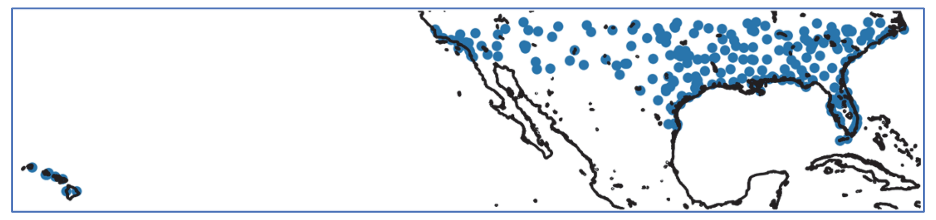
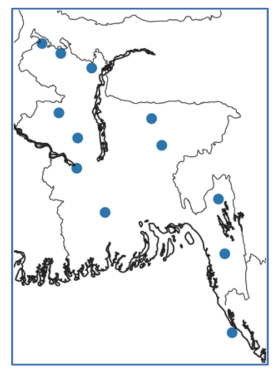
Global rain rate models are crucial in areas ranging from flood management to wireless network planning in developing countries (my interest): *How often does the rain rate exceed x%?*

The current ITU-R model [1] is based on complex model-fitting methods and infrequently updated by the expert working group. An alternative is to use rain rate curves as ***functional data*** [2]:

- We have complete (0% to 100%) "ground truth" rain gauge cdfs at a few hundred points at high time resolution (1-min).
- We have incomplete (0% to ~98%) satellite precipitation radar cdfs with quasi-global coverage.
- We wish to predict rain gauge curves at previously-unseen sites, querying a model learned from rain gauge curves using the radar curves.

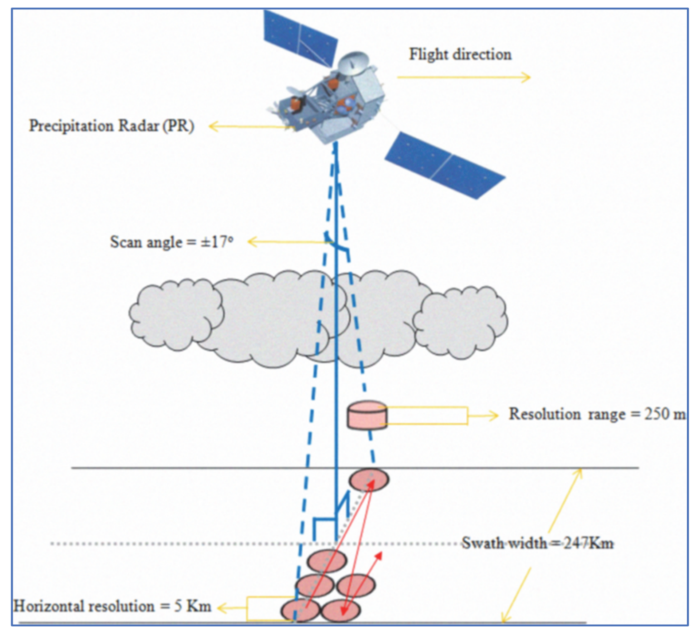
Datasets used

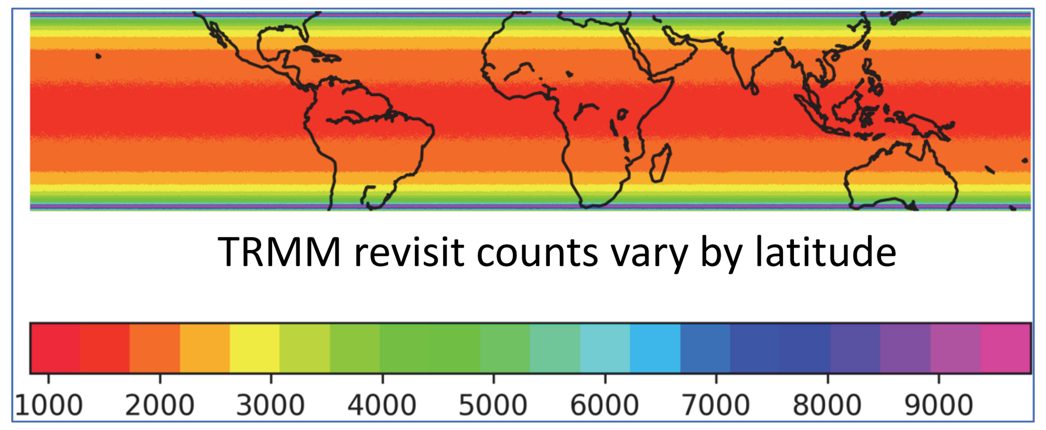
- Inputs for construction of ITU-R (benchmark model) data:
  - Mean surface temperature (from ERA-Interim reanalysis)
  - Mean monthly rain rate (from GPCC)
- Inputs for construction of train/test data:
  - Training set and test set (answer): 1-min resolution rain gauge time series from 308 automated weather stations in Australia, USA, Bangladesh:

Australia is "more like" US data

  - Test set (query): We processed ~44 billion precipitation radar observations (5-km wide radar returns) from the TRMM low-Earth orbit (LEO) satellite mission (1997-2015).



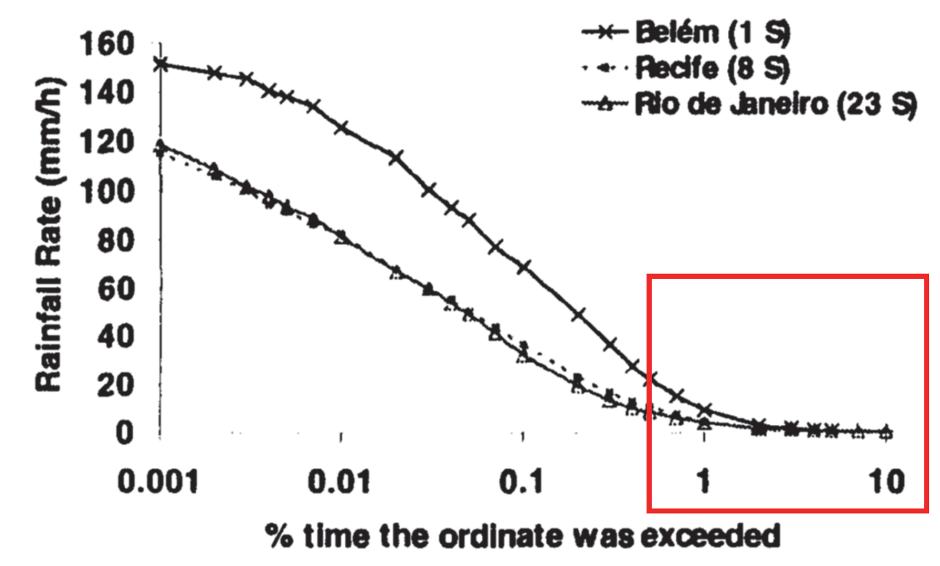


TRMM revisit counts vary by latitude

TRMM radar scan pattern and non-rectilinear ground footprint (left bottom) and non-geostationary orbit (left top) requires "replaying" the orbits using geodetic "hit detection" to recover observations at each site

Features / approach

Complementary cdfs from 0% to 5% (exceedance %) are extracted from gauge and radar data. We learn a model from USA gauge curves to predict the "left" (say, 0% to 2%) part of a curve from the "right." We then query the model using the radar curves which are available to ~2%.



A ccdf from the ITU-R DBSG3 database.

Models

Baseline (ITU): This is computed using a bespoke implementation of the ITU-standard log-normal rain model [1].

Non-parametric functional regression: We apply the kNN-based functional regression scheme of Ferraty & View [3], much as was done in Ciollaro et al. [4] for quasar spectra (intensity) curves:

$$\widehat{f_{left}}(p) = \frac{\sum_{i \in knn(f_{right})} ker(\|f_{right}^{(i)} - f_{right}\|_2 / h) \cdot f_{left}(p)}{\sum_{i \in knn(f_{right})} ker(\|f_{right}^{(i)} - f_{right}\|_2 / h)}$$

where $h = max \|f_{right}^{(i)} - f_{right}\|_2$.

Curve pre-smoothing: Ciollaro et al. pre-smoothed their raw observations, which is common in FDA (but not universal); we chose to use the raw ccdfs as they contain measurement artifacts that should be properly corrected.

Curve amplitude and phase registration: It is also common in FDA to "shift" curves up/down (amplitude) and left/right (phase) prior to analysis; we mean-shift (as a form of instrument bias adjustment) but do not phase-shift (or scale).

Parameter cross-validation: The usual choices for this kind of scheme must all be cross-validated: **kernel choice** (e.g., linear vs. quadratic), the kNN **bandwidth parameter** *k*, etc. Based on 10x 10-fold cross-validation, we chose k=3 and linear kernels.

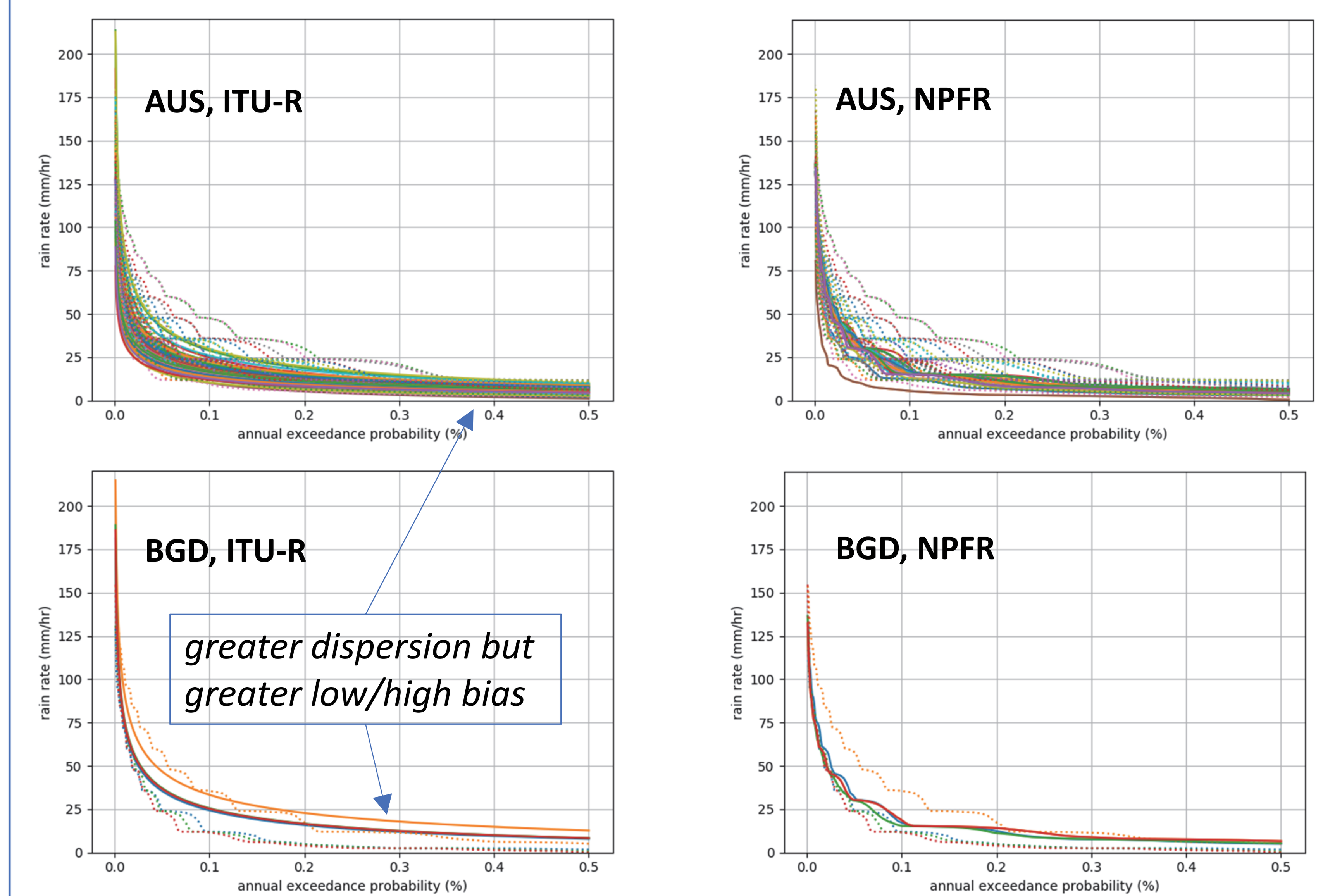
.

Results

ITU-R P.311 standardizes use of mean, std. dev., and rms values of **relative error** (vs. "ground truth") and presentation of results computed at 0.001%, 0.002%, 0.003%, 0.005%, …, 1%, 2%, 3%, 5%.

Predicting the entire 0%-5% ccdf and comparing at these points,

		Mean (%)	std. dev. (%)	rms (%)
AUS (n=45)	ITU-R	-6.38	0.2839	8.35
	NPFR	3.43	0.2930	8.44
BGD (n=4)	ITU-R	52.70	102.22	132.26
	NPFR	24.57	67.89	52.13



(Dashed lines are gauge data, solid lines are predictions.)

Discussion / interpretation

While "expected" (the curves are simple), it's still surprising when a strictly data-driven model performs as well as a parametric model hand-tuned for 25+ years. The data-driven model has *no physical explanatory power*, but that is not always as important as accuracy and rapid updateability for non-geophysics applications.

Future

Future work includes (1) comparison against additional alternatives (e.g, GP FDA) and (2) full benchmarking against the sites in the ITU-R DBSG3 database.

References

1. ITU-R WG3. "Characteristics of precipitation for propagation modelling," Recommendation P.837-7 (2017).
2. Ferraty & Vieu. *Nonparametric Functional Data Analysis* (2006).
3. Ferraty et al. "Regression when both response and predictor are functions," *J. Multivariate Analysis* 109 (2012) 10-28.
4. Ciollario et al. Functional regression for quasar spectra. arXiv:1404.3168 (2014).