Functional Data Analysis for Rain Rate Statistics

Paul M. Aoki <pmaoki> [1]

Abstract

Detailed rain rate frequency statistics ("How many minutes per year does this site receive more than x mm/hr of rain?") have global applications but are only actually measured at a small number of sites worldwide. State-of-the-art (SOTA) estimation models are based on curve-fitting by expert working groups and involve years of effort to update. This project explores the use of *functional data analysis* (FDA) as a potential alternative; we suggest that non-parametric functional regression (NPFR) can be combined with satellite radar data to provide similar or better accuracy for suitable applications.

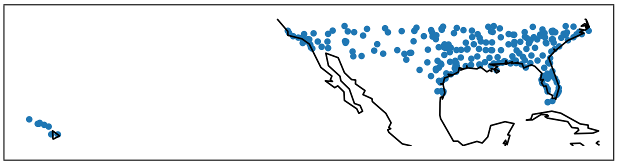

Figure 1. Rain gauge sites: USA (S of 35° N).

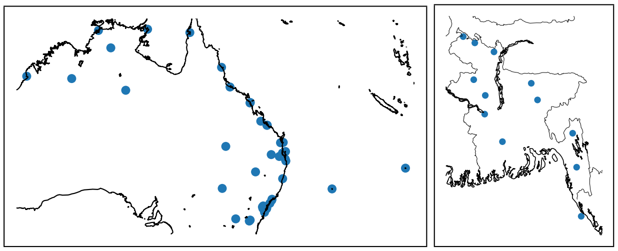

Figure 2. Rain gauge sites: Australia (N of 35° S), Bangladesh.

1. Introduction

Rain rate statistics are important in disciplines ranging from flood management to wireless network design (radio propagation). At any given site, such statistics will likely have to be based on models rather than direct measurement since very few sites have local rain gauges. For example, a radio engineer building a link in rural Bangladesh would consult ITU-R Recommendation P.837-7 (ITU-R, 2017) to estimate the annual *complementary cumulative distribution function* (ccdf) of the rain rate ("What percentage of the year does the rain rate exceed x mm/hr, which will cause my radio link to fail?"). Since 1994, these P.837 models have been laboriously updated using classic deterministic curve-fitting methods with various covariate inputs.[1]

In this project, we examine the use of supervised learning methods – specifically, *functional regression* methods (Ramsay & Silverman, 2005; Ferraty & Vieu, 2012) – to predict annual rain ccdf curves. Specifically, we exploit global-scale remote sensing data that provide partial ccdf curves, querying a functional model trained on full ccdfs using the partial ccdfs.[2] That is, we have a few hundred ccdfs from high-resolution rain gauges that extend from 0% to 100% exceedance, whereas satellite radar gives us global-scale ccdfs that extend from $\sim$ 1% to 100% exceedance. Prediction results so far, while preliminary and unreviewed, have been similar to or better than the SOTA results.

2. Data

While our analysis (i.e., from Section 4 on) will use functional data objects (curves), we extracted the underlying train, test and benchmark ccdfs from several different raw data sets:

Rain gauge data. The "ground truth" data consists of 308 rain gauge time series data, collected at 1-minute resolution over multiple years from sites in the latitude band [35° S, 35° N]. The training set ($n = 259$) is from the USA (Figure 1) with hold-out sets from Australia and Bangladesh (Figure 2). Each time series is subjected to a new implementation of standard NASA rain gauge correction algorithms (Wang et al., 2008) and custom correction/imputation of corrupted records.[3] Per-site ccdfs

[1] CS 229 Fall 2017. Correspondence to: <aoki@acm.org>.

[1] For example, in the P.837-7 model, rain rate is assumed to be "close to lognormal" (Sauvageot, 1994) and have a certain relationship with mean surface temperature.

[2] **CS 229 readers**: this is like the quasar problem in PS#1.

[3] This public data was downloaded from the respective national weather services as part of earlier work (Aoki, 2016) but has been re-processed for this project as the code previously written to do this was proprietary to a previous employer.

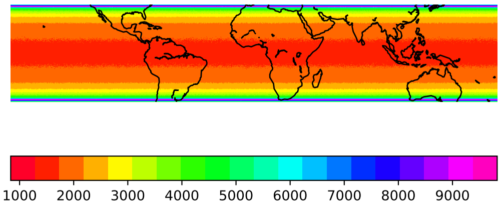

Figure 3. TRMM revisit count varies by latitude.

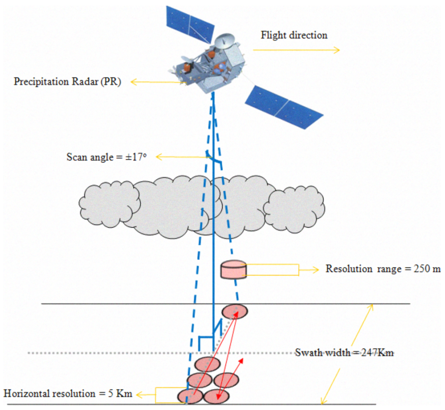

Figure 4. TRMM radar scan pattern (image: JAXA).

were then computed and converted into functional data by sampling as 0.001%-quantiles. Only the (complementary) range 0.001% to 5% were retained (typical annual rain frequencies are 2–3%), so each curve consists of 5000 points.

Satellite radar data. We have obtained time series data (1997-2015) from the TRMM earth observation satellite's precipitation radar (Kozu et al., 2001). TRMM was in a non-Sun-synchronous, low-Earth orbit (LEO) and as such it revisited points on the surface with different frequency depending on latitude (Figure 3). As such, the multi-year time series for any given surface point is much sparser than that of a rain gauge site and we have no exceedance data above $\sim 1\%$.

The raw historical record (over 2 TB) was downloaded in earlier work (Aoki, 2016) but new code was written to extract all radar returns at our rain gauge sites (as opposed to just computing summary statistics). This requires estimating, for each radar "ray" at each orbital track position, the resulting radar footprint on the Earth's surface (the centroid of the $\sim$ 5 km wide pattern traces a sinusoidal path as the radar antenna is scanned from side to side; see Figure 4). Geodetic calculations were computed using GeographicLib (`geographiclib.sourceforge.io`).

Synthetic benchmark data. Using the curve-fitting model of ITU-R P.837-7, we compute synthetic ccdfs for each site. This represents SOTA.

3. Related Work

As previously mentioned, ITU-R P.837-7 is the current SOTA result and is a major improvement over the previous versions of P.837 in terms of accuracy. It remains the most thoroughly-validated and scientifically-justified data set as well.

There have been many alternative curve-fitting models proposed, starting with the Salonen-Baptista model (Salonen & Poiares Baptista, 1997). Several have incorporated TRMM-derived summary statistics (not ccdfs). For example, (Blarzino et al., 2009) used TRMM statistics to de-bias other inputs to the P.837-4 model (resulting in P.837-5), and (Mohd Aris et al., 2013) applied TRMM statistics to replace some inputs to the P.837-6 model (with the aim of increasing accuracy in tropical regions). All use the basic covariate curve-fitting approach and have been superceded by P.837-7 in terms of accuracy.

Rainfall data (in general) are standard examples in any text on FDA (e.g., the "Canadian rain" data in (Ramsay & Silverman, 2005)) but are generally used in direct time series analysis at low time-resolution; a recent example is (Suhaila & Yusop, 2017). FDA has not been applied to analysis of rain rate distributions at high time-resolution, such as 1-minute ccdfs.

Hence, the work reported here is novel in (1) applying FDA techniques to ccdf prediction; (2) use of rain gauge training ccdfs and satellite query ccdfs as functional data; and (3) comparison with P.837-7.

4. Methods

We implemented several alternative methods for extrapolating ccdfs, including two based on conventional functional linear regression (FLR):

(1) Baseline (ITU). Implemented using `numpy` from (ITU-R, 2017; ITU-R WP3J, 2017).

(2) FLR with function-to-function basis splines (FF). Implemented using the R `refund` package in CRAN.

(3) FLR with function-to-function PCA (FFPC). Also implemented using the R `refund` package in CRAN.

(4) Non-parametric functional regression (NPFR). Implemented using `numpy` from (Ferraty et al., 2012).

We rejected (2) and (3) prior to cross-validation. First, these classic functional linear models involve manually selecting basis functions (2) or using basis functions selected via functional PCA (3) (Ramsay & Silverman, 2005). Both can easily result in non-monotonicity, which violates a basic property of ccdfs; see Figure 5 for examples of inappropriate curves. It may be possible to find parameters that

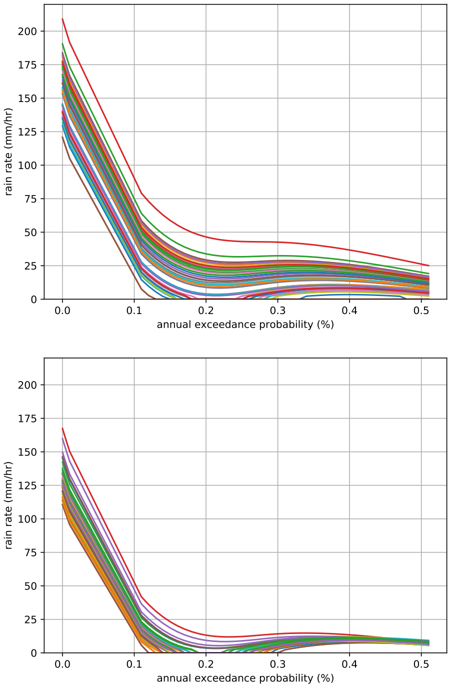

Figure 5. Predicted curves for 45 Australia sites using FF (left) and FFPC with 3 principal component functions (right), both with default basis functions. ("Spaghetti plots" are a standard tool in FDA and are meant to summarize characteristics of an entire set of curves, not individual curves. Here, the reader is just meant to see that the curves are non-monotone.)

produce monotone curves, such a search procedure (like those for curve-fitting) is what we hope to avoid in this project. Second, these methods required decimation of the input data (the FF method consumed over 180 GB of memory on our unmodified Australia data set!).

Hereafter, we compare NPFR methods against the ITU-R P.837 method. We follow the general NPFR approach of (Ciollaro et al., 2014), in which a *functional kernel regression estimator* (Ferraty et al., 2012) is learned from complete functional data examples (quasar spectra) and the learned model is used to predict completed spectra for examples that have only been partially observed due to physical censoring.

4.1. Algorithm choices

Functional regression typically involves several steps[4]:

Pre-smoothing. In FDA, it is common to *smooth* functional data prior to analysis. As in (Ciollaro et al., 2014), we are starting with empirical curves; our ccdf curves do have visible step-like artifacts that result from measurement quantization issues, so we might consider a LOWESS-type smoother (Cleveland, 1979). However, as our curves are already monotone and relatively smooth, we do not.[5]

Registration. In FDA, it is common to *register* the individual functional data prior to analysis (see (Ramsay et al., 2009), Ch. 8). We align the means of the TRMM ccdfs to the rain gauge ccdfs (see (Rice, 2004)), but do not use amplitude (vertical scale) or phase (horizontal translation) registration as our predictions are not invariant to these adjustments.

Kernel estimator. In FDA, many different estimators have been proposed. We follow (Ferraty et al., 2012; Ciollaro et al., 2014) in applying a basic kernel regression estimator based on the Nadaraya-Watson locally-weighted average (see, e.g., (Fan & Gijbels, 1996), §2.2):

$$\widehat{f}_{left}(p) = \frac{\sum\limits_{i \in knn(f_{right})} K\left(\left\| f^{(i)}_{right} - f_{right} \right\|_2 / h\right) f^{(i)}_{left}(p)}{\sum\limits_{i \in knn(f_{right})} K\left(\left\| f^{(i)}_{right} - f_{right} \right\|_2 / h\right)}$$

where $h = \max\limits_{i \in 1,\ldots,m} \left\| f^{(i)}_{right} - f_{right} \right\|_2$. This estimator is not optimal[6] but it is very convenient in that we avoid ccdf validity (monotonicity) issues. In effect, we follow the convention of (Kneip & Utikal, 2001) in modeling such functions as "mixtures" of other functions, albeit for ccdfs.

Kernel. In NPFR, there are several well-established kernels such as the asymmetric triangle kernel[7] or the asymmetric quadratic kernel (as used in (Ferraty et al., 2012; Ciollaro et al., 2014)):

$$K(u) = 2 \cdot (1 - u) \cdot \mathbb{1}_{[0,1]}(u) \qquad \text{(asymmetric triangle)}$$

$$K(u) = \frac{3}{2} \cdot (1 - u^2) \cdot \mathbb{1}_{[0,1]}(u) \qquad \text{(asymmetric quadratic)}$$

4.2. Parameter cross-validation

For the choices not made for qualitative reasons, we selected parameters based on 10-fold cross-validation on the USA (training) rain gauge data, repeated 10 times to reduce noise (see, e.g., (Kuhn & Johnson, 2009) §4.4).

[4] **CS 229 readers**: these were not done in PS#1.

[5] The smoothed quantile functions would also need to be monotone, which is possible if one uses a monotone smoother as in (Ramsay et al., 2009), §5.4.2. However, more importantly, we are not just smoothing out the "usual" measurement noise here, but a systematic artifact of mechanical measurement; a smoother for this problem ought to be a correction and have a justifiable physical basis. This seems beyond the scope of this project.

[6] As is often pointed out (e.g., in (Demongeot et al., 2017)), the asymptotic bias of the basic estimator is higher than that of other local polynomial estimators such as local linear estimators (see (Fan & Gijbels, 1996), §2.3). Others have proposed even more exotic estimators based on block operator matrices (Kadri et al., 2016).

[7] **CS 229 readers**: as used in PS#1.

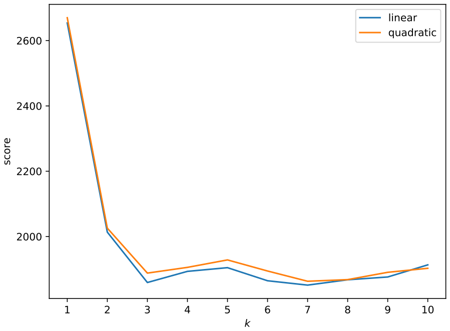

Figure 6. Example cross-validation curve for bandwidth k.

- Kernel choice. We found little difference between the triangle and quadratic kernels; we chose triangle.
- Bandwidth parameter. k generally ranged from 3 to 5 (see, e.g., Figure 6); we chose 3.

5. Experiments

As discussed in Section 4, FLR methods (2) and (3) were rejected due to non-monotonicity; only methods (1) and (4) will be discussed here.

5.1. Quantitative analysis: relative error

We use *relative* error (in %) for site i and (complementary) probability j as our error figure as this is the ITU-specified error figure for P.837 (ITU-R WP3M, 2016; ITU-R WP3J, 2017):

$$\epsilon_j^{(i)} = \frac{\left(\widehat{f}_{left}^{(i)}\right)_j - \left(f_{left}^{(i)}\right)_j}{\left(f_{left}^{(i)}\right)_j} \cdot 100$$

and mean, standard deviation, and weighted rms for this error figure as further specified in (ITU-R, 2015):

$$rms = \sqrt{\frac{\sum\limits_{i=1}^{M} \sum\limits_{j \in 0.001, \ldots, 5} \alpha_i \cdot \left(\epsilon_j^{(i)}\right)^2}{\sum\limits_{i=1}^{M} n_i \cdot \alpha_i}}$$

(where we are comparing at $n_i = 16$ percent-probability levels $0.001, 0.002, 0.003, 0.005, \ldots, 1, 2, 3, 5$ and weighting by site i's time series length α_i in years).

Results on our two hold-out sets are summarized in Table 1. We can see that the performance on all measures are nearly identical for Australia; this makes sense as both Australia and the USA include similar climate variation (from desert to sub-tropics) and both nations have a fair number of rain gauge stations. By contrast, both models do much worse on Bangladesh, which has a monsoon climate unlike the USA and Australia. However, NPFR does notably better, suggesting that the ITU curve-fitting model is more aggressive in its predictions than is warranted (high bias).

In summary, we consider these preliminary results quite promising. While the NPFR model here has no explanatory power (being entirely data-driven), it appears to be able to do as well or better as the expert-tuned curve-fitting model (which represents SOTA) using a fast regression method, limited trainig data, and very basic cross-validation for tuning.

5.2. Qualitative analysis: bias/variance

We can use spaghetti plots to add intuition to the results in Table 1. The main idea of this FDA visualization technique is not to understand individual curves deeply, but rather to see that both models are producing reasonably realistic curves (i.e., predicted curves within the “envelope” of the rain gauge data) overall. We have focused here on the percent-probability range $0.001, \ldots, 0.500$, which is the extreme end of the 1% f_{left} section.

In Figure 7, we have plotted $n = 45$ predicted curves for Australia (solid lines) along with the “ground truth” rain gauge curves at the same sites (dotted lines). The y values are the rain rate (mm/hr) exceeded $x\%$ of the year. As hoped, both models’ curves are plausible. Further, we see that the ITU predictions (top) have an overall low bias (which is confirmed in Table 1).

In Figure 8, we have again plotted $n = 4$ predicted curves for Bangladesh (solid lines) and the corresponding rain gauge curves (dotted lines). Again, even though monsoon climate is poorly represented in the training data (this is true in a different sense for ITU as well), both models produce curves that are plausible. In this case, we see that ITU has an overall high bias (which is again confirmed in Table 1).

Less apparent from Table 1, but visible in both figures, is that NPFR has a somewhat unrealistic lack of dispersion compared to both ITU and the rain gauge curves. This is an area of improvement.

		mean (%)	std. dev. (%)	rms (%)
AUS	ITU	-6.38	0.28	8.35
	NPFR	3.43	0.29	8.44
BGD	ITU	52.70	102.22	132.26
	NPFR	24.57	67.89	52.13

Table 1. Hold-out relative error summaries.

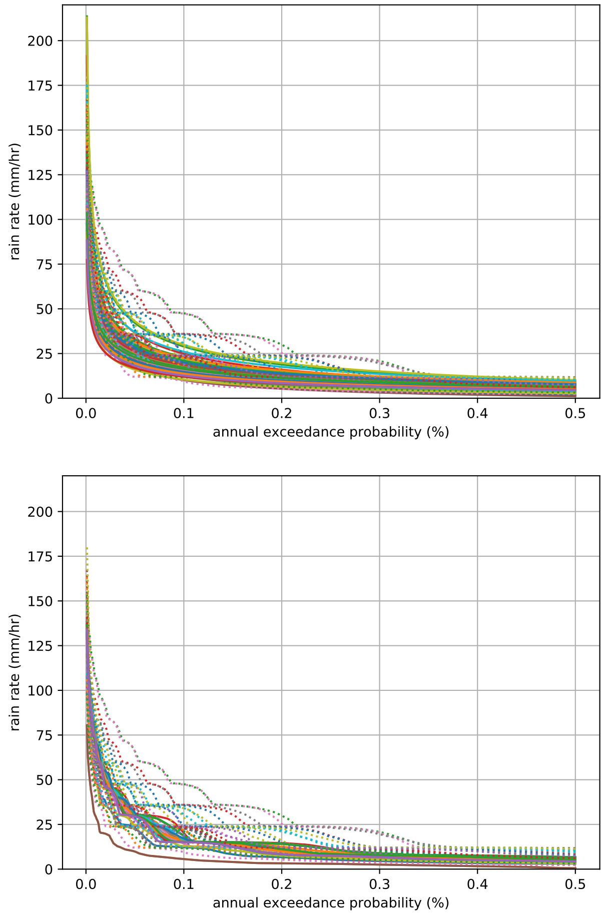

Figure 7. Spaghetti plots, Australia: ITU (top) vs. NPFR (bottom).

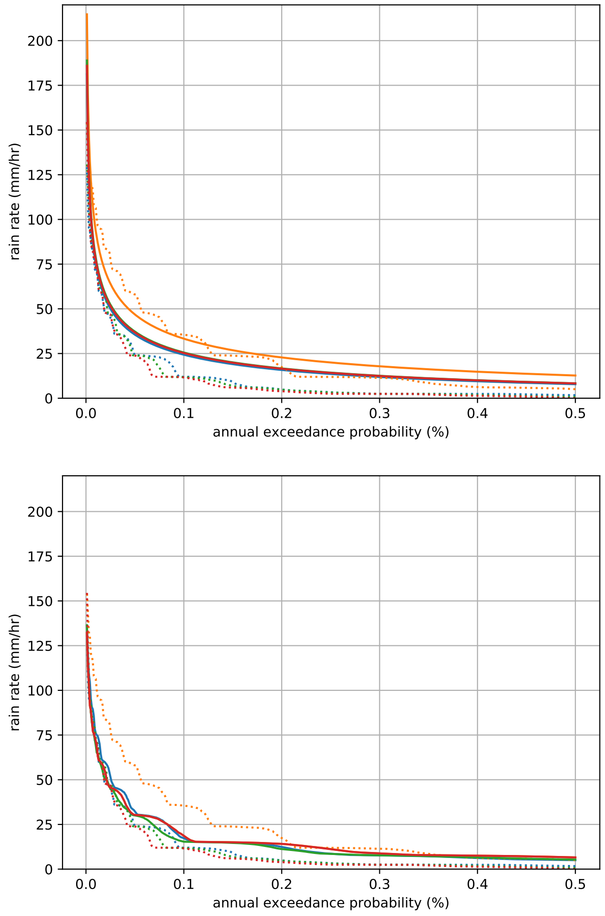

Figure 8. Spaghetti plots, Bangladesh: ITU (top) vs. NPFR (bottom).

6. Conclusions

This project considers three main novel research issues. First, we have examined the use of FDA methods to predict annual exceedance rate curves (ccdfs) for rainfall. Second, we have applied non-parametric functional regression methods to predict complete curves using incomplete query curves derived from satellite radar data. Since these incomplete query curves are available at global scale, this enables direct ccdf prediction where no rain gauge data is available. Third, we suggest that a data-driven approach may be comparable to an expert-tuned model such as P.837-7.

Aside from increased rigor in experimentation (e.g., comparison with additional algorithms), there are many obvious areas for further research:

- The rain gauge data sets used here were freely/cheaply available. There are additional data sets that are not free (e.g., Malaysia) but that will better represent tropical/monsoon climates. Hence, additional data will be critical to improving performance in many areas in developing countries where wireless networks are most needed.
- The TRMM satellite covered only the latitude band [35° S, 35° N]. The follow-on, GPM, has coverage in [60° S, 60° N] but a shorter historical record. A way to combine the two (such that cdfs could be computed, not just summary statistics) would be very valuable.
- Error analysis techniques have been developed for both conventional and non-parametric functional regression methods (the latter based on the bootstrap, as in (Ciollaro et al., 2014)). These are not clearly applicable to the ccdfs here. Some theoretical work may be required to

While not all applications or scientific areas will be satisfied by an approach without explanatory power, this approach should still be useful in commercial engineering applications (such as wireless planning) that prioritize accuracy and updateability.

References

Aoki, Paul M. New rain rate statistics for emerging regions: Implications for wireless backhaul planning. arXiv: 1609.00426, 2016.

Aoki, Paul M. CNNs for precipitation estimation from geostationary satellite imagery. CS 231N project, 2017. URL http://cs231n.stanford.edu/reports/2017/pdfs/557.pdf.

Blarzino, G., Castanet, L., Luini, L., Capsoni, C., and Martellucci, A. Development of a new global rainfall rate model based on ERA40, TRMM, GPCC and GPCP products. In *Proc. EUCAP*, pp. 671–675, 2009.

Ciollaro, Mattia, Cisewski, Jessi, Freeman, Peter E., Genovese, Christopher R., Lei, Jing, O'Connell, Ross, and Wasserman, Larry. Functional regression for quasar spectra. Technical Report arXiv:1404.3168, 2014.

Cleveland, William S. Robust locally weighted regression and smoothing scatterplots. *JASA*, 79:829–836, 1979.

Demongeot, Jacques, Naceri, Amina, Laksaci, Ali, and Rachdi, Mustapha. Local linear regression modelization when all variables are curves. *Statist. Probab. Lett.*, 121: 37–44, 2017.

Fan, J. and Gijbels, I. *Local Polynomial Modelling and its Applications*. Chapman & Hall, 1996.

Ferraty, F., Keilegom, I. Van, and Vieu, P. Regression when both response and predictor are functions. *J. Multivariate Anal.*, 109:10–28, 2012.

Ferraty, Frédéric and Vieu, Philippe. *Nonparametric Functional Data Analysis*. Springer, 2012.

ITU-R. Acquisition, presentation and analysis of data in studies of radiowave propagation. Rec. P.311-15, ITU, 2015.

ITU-R. Characteristics of precipitation for propagation modelling. Rec. P.837-7, ITU, 2017.

ITU-R WP3J. Concerning the rainfall rate model given in annex 1 to recommendation ITU-R P.837-7. Fasc. 3J/FAS/3-E (Rev. 1), ITU, 2017.

ITU-R WP3M. On testing variables used for the selection of prediction methods. Fasc. 3M/FAS/1-E (Rev. 1), ITU, 2016.

Kadri, Hachem, Duflos, Emmanuel, Preux, Philippe, Canu, Stéphane, Rakotomamonjy, Alain, and Audiffren, Julien. Operator-valued kernels for learning from functional response data. *JMLR*, 17:1–54, 2016.

Kneip, Alois and Utikal, Klaus J. Inference for density families using functional principal component analysis. *JASA*, 96:519–542, 2001.

Kozu, Toshiaki, Kawanishi, Toneo, Kuroiwa, Hiroshi, Kojima, Masahiro, Oikawa, Koki, Kumagai, Hiroshi, Okamoto, Kenichi, Okumura, Minoru, Nakatsuka, Hirotaka, and Nishikawa, Katsuhiko. Development of precipitation radar onboard the Tropical Rainfall Measuring Mission (TRMM) satellite. *IEEE Trans. Geosci. & Remote Sens.*, 39:102–116, 2001.

Kuhn, Max and Johnson, Kjell. *Applied Predictive Modeling*. Springer, 2009.

Mohd Aris, Nor Azlan, Luini, Lorenzo, Din, Jafri, and Lam, Hong Yin. 1-minute integrated rain rate statistics estimated from Tropical Rainfall Measuring Mission data. *IEEE Ant. Wireless Propag. Lett.*, 12:132–135, 2013.

Petersen, Alexander and Müller, Hans-Georg. Functional data analysis for density functions by transformation to a Hilbert space. *Ann. Stat.*, 44:183–218, 2016.

Ramsay, J.O. and Silverman, B.W. *Functional Data Analysis (2nd Ed.)*. Springer, 2005.

Ramsay, J.O., Hooker, Giles, and Graves, Spencer. *Functional Data Analysis with R and MATLAB*. Springer, 2009.

Rice, John A. Functional and longitudinal data analysis: perspectives on smoothing. *Stat. Sinica*, 14:631–647, 2004.

Salonen, E. T. and Poiares Baptista, J. P. V. A new global rainfall rate model. In *Proc. ICAP*, pp. 2.182–185, 1997.

Sauvageot, Henri. The probability density function of rain rate and the estimation of rainfall by area integrals. *J. Appl. Meteorol.*, 33:1255–1262, 1994.

Suhaila, Jamaludin and Yusop, Zulkifli. Spatial and temporal variabilities of rainfall data using functional data analysis. *Theor. Appl. Climatol.*, 129:229–242, 2017.

Wang, Jianxin, Fisher, Brad L., and Wolff, David B. Estimating rain rates from tipping-bucket rain gauge measurements. *J. Atmos. Oceanic Technol.*, 25:25–56, 2008.