Optimal Algorithms for Computing Average Temperatures

S. Foucart[1], **M. Hielsberg**[1], **G. Mullendore**[2], **G. Petrova**[1], **P. Wojtaszczyk**[3]

[1]Department of Mathematics, Texas A&M University, College Station, Texas, USA.
[2]Department of Atmospheric Sciences, University of North Dakota, Grand Forks, North Dakota, USA.
[3]Institute of Mathematics, Polish Academy of Sciences, Warsaw, Poland.

Abstract

Numerical algorithms are presented for computing average global temperature (or other quantities of interest such as average precipitation) from measurements taken at specified locations and times. The algorithms are proven to be in a certain sense optimal. The analysis of the optimal algorithm provides sharp a priori bounds on the error between the computed value and the true average global temperature. These a priori bounds involve a computable *compatibility constant* which assesses the quality of the measurements for the chosen model. The optimal algorithm is constructed by solving a convex minimization problem. It is shown that this solution promotes sparsity and hence utilizes a smaller number of well-chosen data sites than those provided. The algorithm is then applied to canonical data sets for the computation of average temperature and average precipitation over given regions and given time intervals. A comparison is provided between the proposed algorithms and existing methods using standard data sets.

1 Introduction

Computing average temperatures is a particular instance of a common task in data processing, namely that of exploiting measurements made on a function f to estimate a quantity $Q(f)$ that depends on f, referred to as Quantity of Interest (QoI) below. In the present situation, $f = f(x,t)$ represents the temperature (in degrees Celsius) as a function of the position x on the surface of the earth and of the time t. We single out temperature as a running example, but other atmospheric features such as humidity or precipitation can be treated equally well. As a QoI, we single out average temperatures over the whole earth or a smaller region R during a whole year or a shorter period Δ, obtained as the normalized integral

$$Q(f) = \frac{1}{\sigma(R)|\Delta|} \int_R \int_\Delta f(x,t)\, d\sigma(x)\, dt, \tag{1}$$

where $\sigma(R)$ is the surface area of the region and $|\Delta|$ is the duration of the period. In going further, we denote by $\Omega := \{(x,t) : x \in R,\ t \in \Delta\}$ the domain of f, so the QoI also reads $Q(f) = |\Omega|^{-1} \int_\Omega f$, where $|\Omega|$ denotes the measure of Ω. Other QoIs such as the temperature $Q(f) = f(\xi,\tau)$ at a specific location ξ and a specific time τ can be considered as well.

Temperatures are continuously monitored by numerous weather stations around the globe, providing us with ample data to estimate the QoI. This data takes the form of a vector

$$w = [f(x_1,t_{1,1}),\dots,f(x_1,t_{1,\ell_1}),\dots,f(x_m,t_{m,1}),\dots f(x_m,t_{m,\ell_m})] \tag{2}$$

whose entries are obtained by recording the temperature at locations x_j and at times $t_{j,i}$, $j = 1,\dots,m$ and $i = 1,\dots,\ell_j$. This vector depends linearly on the temperature function f and we can write it succinctly as $w = M(f)$, where M represents the measurement mapping associated with the data sites $(x_j,t_{j,i})$, $j = 1,\dots,m$, $i = j,\dots,\ell_j$.

This data alone is not sufficient to guarantee that a QoI such as annual global temperature can be accurately computed. Indeed, without additional knowledge, the temperature function could for example oscillate wildly between the data sites. What forbids such unrealistic scenarios are additional

---

Important note: This is the original version. It differs from the revised version.

properties of the temperature function, which are usually described by structural assumptions on f. Such assumptions are referred to as model class assumptions. They quantify the uncertainty in our knowledge of f. Model classes will be discussed in more detail below.

For now, given the data and the model class, we focus on the following three questions:

(i) Is there an optimal accuracy at which one can estimate a QoI?
(ii) Are there procedures that achieve this accuracy?
(iii) Can one implement such procedures?

These questions form the cornerstone of a mathematical theory called *optimal recovery* (see *Micchelli and Rivlin* [1985]). Regarding (ii), for instance, the theory guarantees that there is an optimal algorithm for computing an average temperature $Q(f)$ that takes the form of a weighted average of the data, i.e., making the approximation

$$Q(f) \approx \sum_{j=1}^{m} \sum_{i=1}^{\ell_j} a_{j,i}^* f(x_j, t_{j,i}). \tag{3}$$

Note that current methods used to estimate average temperatures, such as *Hansen and Lebedeff* [1987], *Hansen, Ruedy, Glascoe, and Sato* [1999], and *Hansen, Ruedy, Sato, and Lo* [2010], although not explicitly stated as such, also take the form of a weighted average. However, the weights used in these methods are derived in a rather *ad hoc* manner and are not optimally chosen. Thanks to the introduction of a new and realistic family of model classes for the underlying function f, significant advances were recently made on (i)-(ii)-(iii) in *DeVore, Foucart, Petrova, and Wojtaszczyk* [20xx], where it was revealed how to explicitly construct optimal algorithms for the estimation of QoIs relative to these new model classes. The purpose of this letter is to make geophysical scientists aware of these developments in optimal recovery, so that they can capitalize on their domain expertise to put forward the most realistic model classes for global temperature and thereby utilize the above theory to design optimal algorithms with certifiable performance.

2 Methods

We now formalize our approach, in particular by clarifying the notion of optimal algorithms. By an algorithm, we simply mean a mapping taking a measurement vector $w = M(f) \in \mathbb{R}^m$ as input and returning a number $A(w) \in \mathbb{R}$ as an output, hopefully close to the true QoI $Q(f)$. The error made by this approximation is

$$E(w, A) := |Q(f) - A(w)|, \qquad w = M(f). \tag{4}$$

The performance of the algorithm A is then assessed in a worst-case setting, keeping in mind that, besides the data, the only information one has (or assumes) about f is that f comes from the model, that is f belongs to a class K of functions referred to as a *model class*. Thus, we introduce

$$E(K, A) := \sup_{f \in K} E(M(f), A) \tag{5}$$

as an indicator of the performance of the algorithm over the class K. A favorable bound on $E(K, A)$ guarantees that $Q(f)$ is computed well. Finally, an optimal algorithm A^* is one that makes $E(K, A)$ as small as possible, i.e.,

$$E(K, A^*) = \inf_{A} E(K, A) =: E^*(K). \tag{6}$$

Note that $E^*(K)$ quantifies the optimal performance relative to the class K.

The framework presented above is of course very much dependent on the model class K. As an ideal situation, we would prefer to deduce properties of the temperature function f from first principles of atmospheric science, say from a system of partial differential equations (PDEs) governing the behavior of f. Such properties could, for example, be inferred by regularity theorems for PDEs, and may quantify the smoothness of f. Unfortunately, in the case of the temperature function, the complexity of the system of PDEs makes regularity theorems hard to exploit. Therefore, the model classes we advocate are not built on smoothness but instead on approximability.

We start from the observation that, as in quadrature theory, estimation procedures are usually constructed through approximation, i.e., one selects a method to approximate f, e.g. by polynomials, piecewise polynomials, radial basis functions, etc., and then designs an algorithm based on this approximation. By choosing a specific form of approximation, there is an implicit acceptance that f is well approximated by the selected method. This tacitly suggests to consider a model class defined as

$$K = K(\epsilon, V) := \{g \in C(\Omega) : \ \mathrm{dist}(g, V) \leq \epsilon\} \tag{7}$$

involving an approximation threshold $\epsilon > 0$ and a set V of functions used as approximants (in the context of this paper, V is a linear space of finite dimension n). The distance is relative to the usual norm on the space $C(\Omega)$ of continuous functions, so that

$$\mathrm{dist}(g, V) = \inf_{v \in V} \max_{\omega \in \Omega} |g(\omega) - v(\omega)|. \tag{8}$$

Regarding (i), it was shown in *DeVore, Foucart, Petrova, and Wojtaszczyk* [20xx] that the optimal performance can be precisely evaluated for the model class (7). It decouples as

$$E^*(K(\epsilon, V)) = \mu_V \epsilon, \tag{9}$$

revealing that, besides the approximation capability of V, another important part is played by a constant μ_V that encapsulates the compatibility of the data sites $(x_j, t_{j,i})$, $j = 1, \ldots, m$, $i = 1, \ldots, \ell_j$, with the space V. A poor compatibility will result in a large μ_V [1]. The roles of μ_V and ϵ are somewhat competing in the choice of a proper linear space V on which to build the algorithm. Indeed, we want a space with both good approximation capability, so that ϵ is small, and good compatibility with the data, so that μ_V is small. But enlarging the space V will have the effect of decreasing ϵ while increasing μ_V. Such a tension is similar to the one encountered in statistical learning when confronted with the problem of overfitting the data.

Regarding (ii) and (iii), *DeVore, Foucart, Petrova, and Wojtaszczyk* [20xx] also put forward the construction of an optimal algorithm A^* for the estimation of real-valued linear QoIs such as (1). For easy comparison with other algorithms, we describe this construction in the special framework where there is no time dependence, i.e., when $f = f(x)$ depends only on the position x. Existing algorithms for estimating annual temperatures can indeed be unscrambled as producing beforehand an annual temperature $f(x_j)$ at each data site x_j, and then computing a weighted average of the $f(x_j)$. We place ourselves in the same setting where data consist of measurements $f(x_j)$ which have already averaged the time. Another instance of the time-independent framework is the case where $f_\tau(x) = f(x, \tau)$ represents an instantaneous temperature at a given time τ.

In the time-independent framework, given data sites $(x_1, \ldots, x_m)$ and a basis $(v_1, \ldots, v_n)$ for the space V, the optimal algorithm A^* for the estimation of a real-valued linear QoI Q first produces a solution

$$a^* := \mathrm{argmin} \left\{ \sum_{j=1}^m |a_j| : \ \sum_{j=1}^m a_j v_i(x_j) = Q(v_i), \quad i = 1, \ldots, n \right\} \tag{10}$$

of a ℓ_1-minimization problem with variable $a = [a_1, \ldots, a_m]$, which can be reformulated as a linear optimization problem and solved using standard techniques. Notice that the optimal weights $a_1^*, \ldots, a_m^*$ depend on the sites $x_1, \ldots, x_m$ but not the data $f(x_1), \ldots, f(x_m)$ at these sites. Next, for each measurement vector $w \in \mathbb{R}^m$, the algorithm computes the weighted average

$$A^*(w) := \sum_{j=1}^m a_j^* w_j. \tag{11}$$

The algorithm A^* is optimal in the sense that $E(K(\epsilon, V), A^*) = E^*(K(\epsilon, V))$ for all $\epsilon > 0$. Notice that the knowledge of $\epsilon > 0$ is not necessary to construct the algorithm A^*. The compatibility constant

[1] The precise definition of μ_V would reveal that the extreme situation $\mu_V = +\infty$ occurs if V contains a nonzero function v such that $v(x_j, t_{j,i}) = 0$, $j = 1, \ldots, m$, $i = 1, \ldots, \ell_j$, and this automatically happens when $\ell_1 + \cdots + \ell_m < n$.

μ_V is incidentally obtained as a byproduct of (10), since

$$\mu_V = 1 + \sum_{j=1}^{m} |a_j^*|. \tag{12}$$

The algorithm A^* possesses several pleasing attributes that are worth pointing out. As noted, the minimization procedure (10) does not involve the data but only the locations. Therefore, the algorithm produces offline the weights once and for all and the formula (11) can be reused instantly for each data $w_{\tau'} = [f_{\tau'}(x_1), \ldots, f_{\tau'}(x_m)]$ recorded at another time τ' (or other time averages). The only proviso here is that the same weather stations must provide measurements at the same times (or averages over the same time period). Note that the optimal algorithm generally selects only a small number of station locations. This is due to the model assumption. The merit of this observation lies less in terms of data collection (data from all stations should indeed be kept, since the selection made by the algorithm depends on the available stations, the space V, and the QoI to be estimated) but more in terms of data transmission: massive databases are commonly stored in a cloud system accessible to multiple scientific teams, and downloading data to estimate a QoI can be a bottleneck, so it is clearly beneficial to download only the portion that will be used in the computation.

3 Experimental results

We now test our method on several situations encountered in atmospheric science: the estimation of average annual global temperature, the estimation of average seasonal regional temperature, and the estimation of total annual global precipitation.

3.1 Average annual global temperatures

Our first and most exhaustive experiment illustrates the application of our algorithm to the estimation of average annual global temperatures. The results are compared with the ones released by agencies such as NOAA National Climatic Data Center and NASA Goddard Institute for Space Studies. In these cases, the time dependence is discarded by considering for each year a function $f = f(x)$ representing an average annual temperature depending only on the position x.

Data provenance: The experiment relies on the following standard data sets:

- Raw Land Data (RLD), obtained by merging monthly land-based station temperatures (GHCNM) from *Lawrimore et al.* [2011] and monthly Antarctic land-based station temperatures (SCAR MET-READER) from *SCAR MET-READER* [2018];
- Processed Land Data (PLD), obtained by processing the RLD using GISTEMP steps 0–2, see *GISTEMP* [2018] and *Hansen, Ruedy, Sato, and Lo* [2010];
- Gridded Land Data (GLD), using the PLD to assign temperatures for each grid center in the covering of the globe by 8,000 equal-area cells, see *Hansen and Lebedeff* [1987];
- Gridded Sea Data (GSD)[2], downloaded from the GISTEMP's website https://data.giss.nasa.gov/pub/gistemp/SBBX.ERSSTv4.gz and obtained following *Huang et al.* [2015] and *NOAA National Centers for Environmental Information* [2018].

Algorithmic details: Our numerical algorithms require the choice of an approximation space V. We consider spherical harmonics of degrees $L = 3$, $L = 6$, and $L = 9$, which we denote by SH3, SH6, and SH9. We also consider approximation by piecewise constant functions on two standard partitions of the globe, namely the coarse and fine partitions used in *Hansen and Lebedeff* [1987], *Hansen, Ruedy, Glascoe, and Sato* [1999], *Hansen, Ruedy, Sato, and Lo* [2010]. We denote the corresponding optimal algorithms by PCC and PCF. We compare the performance of our algorithms with two standard methods provided by GISTEMP and by *NOAA National Centers for Environmental*

[2] These data have undergone some processing and gridding steps, too, but we were unable to obtain the raw data used to generate them .

Information [2018]. Both of these methods implicitly rely on piecewise constant approximation, however, they differ from the optimal algorithm for approximation model. We compare these algorithms on two data sets:

Data set 1: This data set consists of the merging of GLD and GSD. This is the standard data set used in GISTEMP.

Data set 2: This data set consists of the merging of PLD and GSD. This set works more closely with raw data, at least on land.

Observations: The results using PCC and PCF as well as using SH3, SH6 and SH9 on both data sets 1 and 2 are displayed in Figures 1 and 2. In these figures, temperature anomalies in $^{\circ}C$ were computed from the 1951–1980 baseline average. They are compared with those computed using GISTEMP and those reported by NOAA over the 1950–2016 time period. The results call for a number of comments:

- Our results match those reported by NOAA and NASA tightly for data set 1, because they rely on the exact same data set, and more loosely for data set 2, which is closer to raw data;
- Spherical harmonics are surprisingly effective, considering that $\dim(V) = (L+1)^2 = 100$ for $L = 9$ while $\dim(V) = 8{,}000$ for the fine grid;
- Moving from the coarse to the fine grid improves the approximation capability ϵ and does not severely deteriorate the compatibility constant μ_V, and so produces more accurate curves;
- Likewise, for spherical harmonics, increasing L improves ϵ and does not severely deteriorate μ_V. However, if the stations were not reasonably spread around the globe (for example by considering land stations only), then μ_V could increase and hence impact the QoI estimation;
- The values of μ_V are very close to the least possible value of 2 throughout for all the experiments from Figures 1 and 2.

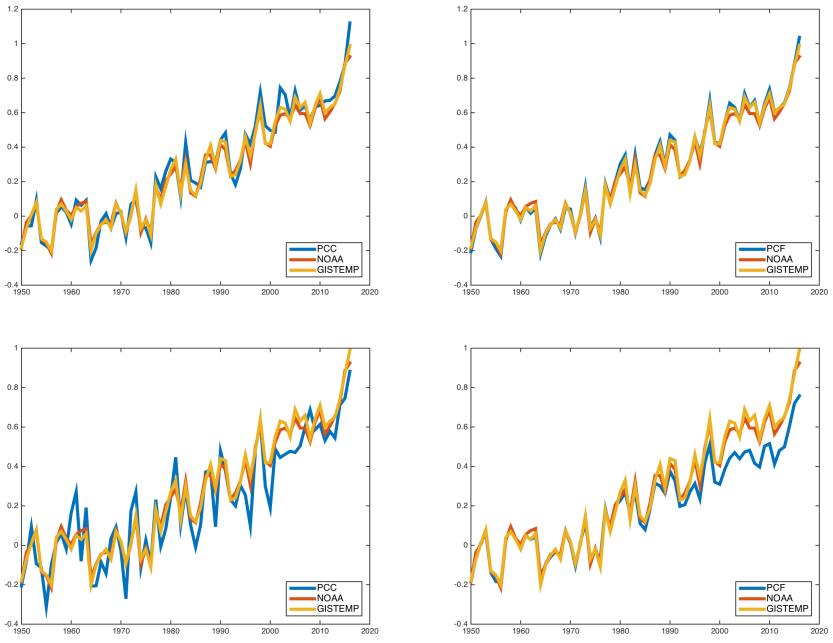

Figure 1: Temperature anomalies in $^{\circ}C$ computed using PCC (left column) and PCF (right column) with data set 1 (top row) and data set 2 (bottom row). The values reported by NOAA and by GISTEMP are also shown.

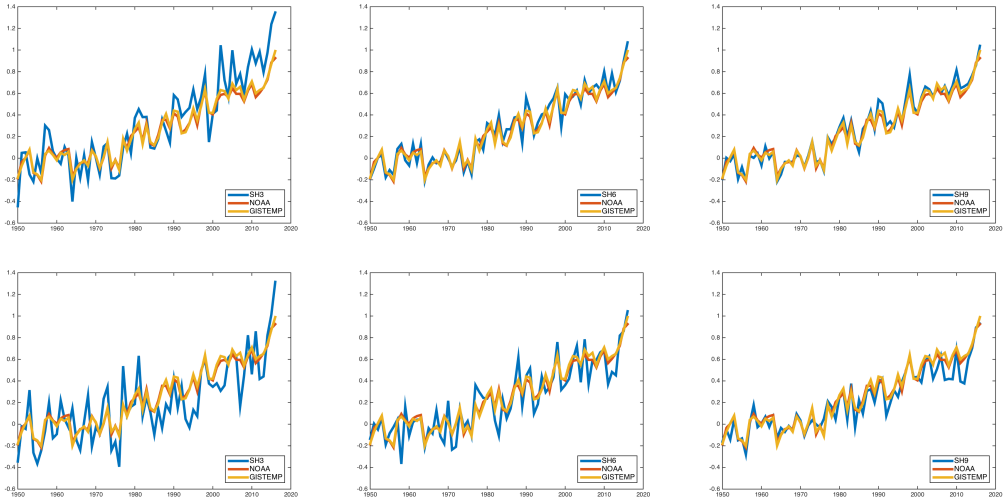

Figure 2: Temperature anomalies in $^\circ C$ computed using SH3 (left column), SH6 (middle column) and SH9 (right column) with data set 1 (top row) and data set 2 (bottom row). The values reported by NOAA and by GISTEMP are also shown.

Weather station positioning: According to properties of ℓ_1-minimizers, the solution a^* of (10) is sparse, meaning that only $n = \dim(V)$ weights among $a_1^*, \ldots, a_m^*$ are nonzero, which implies that only n from the m weather stations are involved in the optimal estimation of average temperatures (Figure 3; recall that the computation of a^* does not depend on the measured temperatures but only on the positions of the stations). In our spherical harmonics experiments, we have observed that the n stations selected by the method tend to be evenly spread and that clustered stations, as in the land-only data set, tend to produce larger values of μ_V , so the QoI estimation becomes less reliable. When L increases, the value of μ_V does not severely deteriorate as long as there are sufficiently many stations which are reasonably spread around the globe.

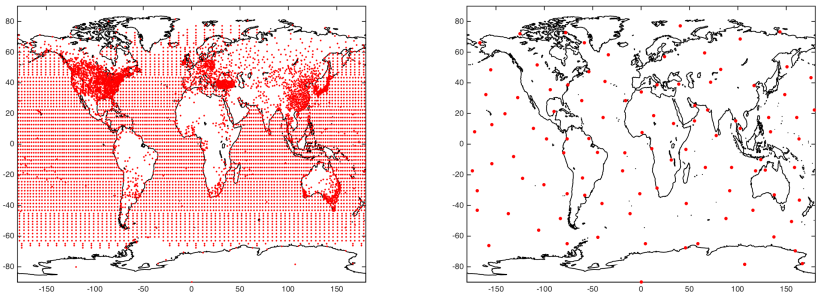

Figure 3: The set of $m = 9{,}187$ locations available in data set 2 (left) along with the $n = 100$ selected locations for SH9 (right) in 1985. The compatibility constant here is $\mu_V = 2$.

3.2 Average seasonal regional temperatures

We apply our method to estimate the average temperature in the state of Texas over two periods of three months (winter and summer) from 2000 to 2016. Time dependence is now incorporated.

Thus, the space V consists of functions of two variables: the position x and the time t. We assume a piecewise constant dependence on x and a piecewise linear dependence on t, with breakpoints every week. The data we used underwent a preprocessing step producing average weekly temperatures at each weather station from daily values acquired from *Menne et al.* [2012]. By contrast to the annual global temperature, we do not need to discard a weather station from the record just because one weekly reading is missing.

Figure 4: The 115 station located in Texas (left) and the temperature anomalies in $^\circ C$ for summer (middle) and winter (right) estimated with the incorporated time-dependence and compared with values reported by NOAA. The volume of data involved in the optimal estimation was reduced by a factor up to $115/19 \approx 6$.

3.3 Total annual global precipitations

For our precipitation computations, we use the Global Precipitation Climatology Centre monthly precipitation data set from *Schneider et al.* [2011]. Intuitively, the task should be difficult because precipitations are much more irregular geographically than temperatures, so finding a good model class could be challenging. Figure 5 shows the results using SH3, SH6, and SH9 and compares them with values reported by *Blunden and Arndt* [2016], downloaded from *EPA: Climate Change Indicators* [2016]. We observe that the spherical harmonics model produces less oscillation. However, we stress that the precipitation data set, which includes only stations over land, does not give full coverage of the globe. This leads to a less accurate QoI estimation due to a larger compatibility constant μ_V.

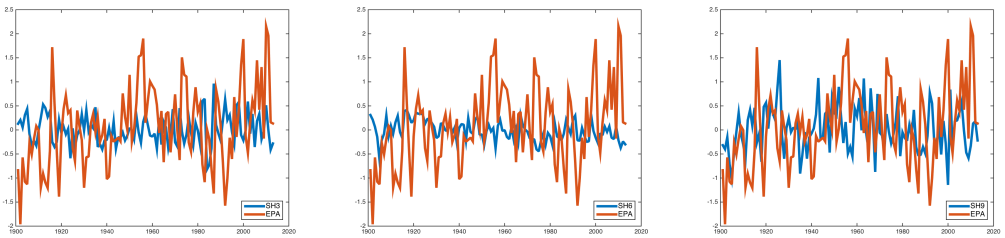

Figure 5: Annual precipitation computed using SH3 (left), SH6 (middle) and SH9 (right). The values reported by *Blunden and Arndt* [2016] are also shown. The values of the compatibility constant are $\mu_V = 2.000002$, $\mu_V = 2.192771$, and $\mu_V = 5.343402$, respectively.

4 Discussion and conclusion

We have acquainted the geophysical community with a template method to estimate diverse QoIs, with a special emphasis on average temperature. The method, whose parameters are computed offline once and for all, is optimal with respect to the approximation model selected. Ideally, the method should be applied to raw data, with data manipulation as performed by other methods (such as urban adjustment) to be integrated in the modeling stage, i.e., the choice of the space V. This choice is of course critical for the reliability of the method. It is rather surprising that spherical harmonics provided decent results — this would have been conceivable for temperatures in the free troposphere, but on the earth surface they obliterate the important dependence of temperature on latitude, altitude, surface type, etc. We hope that domain experts can chime in and propose relevant choices of V for the model. A good space V should have a small dimension (the space of piecewise constants does not) for at least two reasons: the compatibility constant μ_V will remain small and the estimation formula will feature few nonzero weights. The latter is due to the automatic sparsity of ℓ_1-minimizers. It does not suggest eliminating numerous weather stations and acquiring less data, as the nonzero weights depend on the QoI, on V, and on the available stations, but it presents an interesting advantage in terms of data transmission.

Acknowledgments

R. D. partially supported by NSF grant DMS-1521067, ONR N00014-15-1-2181, and ONR N00014-16-1-2706. S. F. partially supported by NSF grant DMS-1622134. G. M. partially support by NSF grant ACI-1450168. P. W. partially supported by National Science Centre, Poland grant UMO-2016/21/B/ST1/00241.

References

Blunden, J. and Arndt, D.S. (eds.). (2016). State of the climate in 2015. *B. Am. Meteorol. Soc.* 97(8):S1-S275.

Climate Change Indicators: US and Global Precipitation. Dataset accessed 2018-04-27 at https://www.epa.gov/sites/production/files/2016-08/precipitation_fig-2.csv.

DeVore, R., Foucart, S., Petrova, G., Wojtaszczyk, P. (20xx). Computing a quantity of interest from observational data. *Constructive Approximation*, To appear.

GISTEMP Team, 2018: GISS Surface Temperature Analysis (GISTEMP). NASA Goddard Institute for Space Studies. Dataset accessed 2018-04-27 at https://data.giss.nasa.gov/gistemp/.

Hansen, J. and Lebedeff, S. (1987). Global trends of measured surface air temperature. *Journal of Geophysical Research* 92.D11, 13.

Hansen, J., Ruedy, R., Glascoe, J., and Sato, M. (1999). GISS analysis of surface temperature change. *Journal of Geophysical Research: Atmospheres* 104, no. D24, 30997–31022.

Hansen, J., Ruedy, R., Sato, M., and Lo, K. Global surface temperature change. *Reviews of Geophysics* 48, no. 4.

Huang, B., Banzon, V. F., Freeman, E., Lawrimore, J., Liu, W., Peterson, T. C., Smith, T. M., Thorne, P. W., Woodruff, S. D., and Zhang, H. (2015). Extended Reconstructed Sea Surface Temperature (ERSST), version 4. *NOAA National Centers for Environmental Information* doi:10.7289/V5KD1VVF.

Lawrimore, J. H., Menne, M. J., Gleason, B. E., Williams, C. N., Wuertz, D. B., Vose, R. S., and Rennie, J. (2011). An overview of the Global Historical Climatology Network monthly mean temperature data set, version 3. *J. Geophys. Res.* 116, D19121.

Menne, M.J., Durre, I., Vose, R.S., Gleason, B.E. and Houston, T.G.. An overview of the Global Historical Climatology Network-Daily Database. *Journal of Atmospheric and Oceanic Technology* 29, 897-910, doi:10.1175/JTECH-D-11-00103.1.

Micchelli, C. and Rivlin, T. (1985). Lectures on optimal recovery. Numerical analysis, (Lancaster, 1984), 21–93, Lecture Notes in Math., 1129, Springer, Berlin.

NOAA National Centers for Environmental Information, Climate at a Glance. *Global Time Series*, published February 2018, retrieved on February 28, 2018 from http://www.ncdc.noaa.gov/cag/

REference Antarctic Data for Environmental Research (READER), Scientific Committee on Antarctic Research, http://www.antarctica.ac.uk/met/READER, Accessed 2018-04.

Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Rudolf, B., Ziese, M. (2011). GPCC Full Data Reanalysis Version 6.0 at 1.0°: Monthly Land-Surface Precipitation from Rain-Gauges built on GTS-based and Historic Data. doi: 10.5676/DWD_GPCC/FD_M_V7_100.