Bias correction in estimation of public health risk attributable to short-term air pollution exposure

Wesley S. Burr\textsuperscript{a}, Glen Takahara\textsuperscript{b} and Hwashin H. Shin\textsuperscript{a,b,*}

Numerous epidemiologic studies have reported associations between short-term air pollution exposure and mortality. Such short-term risk models include smooth functions of time to control for unmeasured confounding variables. We demonstrate bias in these short-term Generalized Additive Model estimates because of lack of accounting for long timescale variations and propose a family of improved time smoothers to reduce and control the bias. The strengths of the proposed smoother are twofold: a clear separating of short-term and long-term effects and an obvious choice of smoothing parameters from pre-determined timescales of interest. We demonstrate improvements through simulations and analysis of examples of air pollution and mortality data from Chicago, Ill. from the National Morbidity, Mortality and Air Pollution Study database, showing reduced bias in the risk estimates. © 2015 The Authors. Environmetrics published by John Wiley & Sons Ltd.

Keywords: air pollution; generalized additive models; semi-parametric models; multitaper spectrum estimation; natural cubic regression splines; mortality; Slepians

1. INTRODUCTION

Numerous studies have shown that short-term exposure to air pollutants is associated with adverse health effects on population, in particular increased risk of premature mortality (Schwartz and Marcus, 1990; Burnett \textit{et al.}, 1999; Dominici \textit{et al.}, 2006; Shin \textit{et al.}, 2012). In many of these studies, the short-term effect of particulate matter (PM) on health has been the focus, with significant effort expended in the model building on accounting for unmeasured confounders (Dockery and Pope, 1994; Schwartz \textit{et al.}, 1996; Dominici \textit{et al.}, 2000; Peng \textit{et al.}, 2006). These unmeasured confounders are considered to be factors that vary with time in a manner similar to both mortality and air pollution.

The short-term adverse health effects of air pollutants are typically estimated using Poisson Generalized Linear Models (GLMs) or Generalized Additive Models (GAMs). The short-term risks are estimated from short timescale variations in mortality and air pollution, protected from confounding by unmeasured longer timescale variations such as demographic changes. This is carried out by including smooth functions of calendar time using a GAM framework (Kelsall \textit{et al.}, 1997; Dominici \textit{et al.}, 2000), typically natural cubic regression splines (Dominici \textit{et al.}, 2002; Ramsay \textit{et al.}, 2003; Peng \textit{et al.}, 2006). The hope is that little information from timescales longer than a pre-specified cutoff will be included in the short-term risk estimate.

In this paper, we demonstrate a new way of examining these models that include smooth functions of time to show that the use of low (6–8) degree-of-freedom (df)/year spline-based smooth functions of time does not completely protect the short-term risk estimate from confounding by long timescale associations. To be clear, we consider long timescale variations to be those with periods in the order of 60 days and longer, as the usual interpretation in the literature of a 6 df/year spline function of time. As we will see in the following sections, while this interpretation is intuitive, it should not be taken at face value.

In Section 2, we briefly review the standard model for estimation of short-term risk because of air pollution and discuss the literature supporting the inclusion of the smooth function of time along with its implications. Section 3 examines semi-parametric GAMs using a simplified risk model. A derivation is provided for the estimator that suggests three possible sources of error due to the time-based smoother. Section 4 motivates and introduces an alternative smooth function of time and the minimization of error. Finally, Section 5 examines a number of simulation studies that demonstrate a clear reduction in bias and provides an example taken from the NMMAPS data database, Peng and Welty (2004).
2. ESTIMATING POPULATION HEALTH RISK

2.1. Standard model and the smooth function of time

Work completed in the 1990s and 2000s (Schwartz and Marcus, 1990; Schwartz et al., 1996; Burnett et al., 1999; Dominici et al., 2000; Dominici et al., 2002; Peng et al., 2006) led to the development and standardization of the current model for the estimation of population health risk due to air pollution. Built around a Poisson GLM or GAM, the logarithm of daily mean mortality counts is modeled as an additive combination of a smooth function of daily mean temperature, at least one smooth function of time (which we will discuss in further detail in the succeeding paragraphs), a categorical variable for the day-of-week effect, and a parametric air pollution covariate such as PM$_{10}$ or Ozone. A generalized version of such a model is presented in Eqn. 1, with a specific working example provided in Eqn. 2. For $Y_i$, a daily mortality count series indexed by day, we write

$$\log(\mu_i) = \beta_0 + x_i \beta_1 + \text{DOW}_i \beta_2 + \sum_{j=1}^{k} f_j (z^j, \lambda_j)$$

(1)

where $\mu_i = E[Y_i]$ is the time-varying daily mean mortality, $x_i$ is the pollutant of interest, again measured via a daily metric, such as sample mean, DOW$_i$ is an indicator function for the day-of-week, and the $f_j (z^j, \lambda_j)$ are $k$ arbitrary smooth functions of environmental covariates $z^j$ with respective degrees-of-freedom $\lambda_j$ (possibly including a time variate). A specific example (Model 2) of such a model that is used as a baseline for the rest of the paper is given as

$$\log(\mu_i) = \beta_0 + x_i \beta_1 + \text{DOW}_i \beta_2 + S(\text{Temp}, df = 3) + S(\text{Time}, df = 6/\text{yr})$$

(2)

where both S functions are natural cubic regression splines with the indicated df, computed on calendar time and daily mean temperature, both on the same day as the response. Non-penalized fixed-df cubic regression splines are used due to the findings of Ramsay et al. (2003) and Dominici et al. (2004), which showed that penalized splines can lead to collinearity (the non-parametric analogue of multicollinearity) for these models.

The smooth function of time is standard and is included to account for unmeasured confounders. Mortality has a strong seasonal variation, which in Canada and the USA is largely driven by the seasons of the continental and subtropical climates. There can be any number of causal factors that contribute to non-accidental mortality, many of which are not measured routinely or easily. In many of these causal risk factors, long timescale structure is also present, for example, influenza epidemic cycles peak in mid-winter for the northern hemisphere, and demographic shifts vary on multi-year or decadal time scales. Accordingly, the standard model of Eqn. 2 includes a smooth function of time to account for these unmeasured risk factors. Additionally, the smooth function of time is included so as to account for any additional serial correlation in the log mortality count series. The intention is to allow for the short-term risk to be estimated using only short-term (temporal) variations in mortality and air pollution by removing long timescale variation.

2.2. Performance of the smooth function of time in the standard model

Traditionally, a diagnostic used to evaluate the performance of the smooth function of time has been a plot of the sample autocorrelations of the residuals. Unfortunately, the autocorrelation does not show sufficient detail so as to be able to detect high-coherence but low-magnitude relationships such as exist in the seasonal band (around 3 months period). To demonstrate this, consider Figure 1, which shows (dark grey dashed-with-circles line) the autocorrelation function of the residuals from the application of Model 2 to the data from Chicago, IL, 1987–2000, using all-cause mortality as response, lag-0 trimmed-mean PM$_{10}$ as pollutant, and a Poisson family link function. The data used came from the NMMAPSdata database (Peng and Welty, 2004), with one modification: the lag-0 trimmed-mean PM$_{10}$ series provided in the database was linearly interpolated (direct line segments joining endpoints of missing gaps) for a small number of missing points in order to demonstrate the power spectrum approach. For full details, see the source code (Burr, 2015).

It is difficult to tell what the interpretation of the dashed (with circles) line should be in Figure 1, except to say that the residual autocorrelation values tend to the negative, and are definitely not white noise. The grey banded 95% confidence region in Figure 1 is exceeded by fully half of the lags shown, when a white noise series would have two or three at most. The first three lags (1, 2, and 3 days) are positive, indicating a day-to-day structure in the residuals that is usually indicative of long-term trends, including seasons and longer-term periodicities, such as demographic change. Typically, the residual serially correlated structure could be accounted for, by including additional functions of time in the model, for example, smooth functions of time with 1 df/year (to account for demographic shift). Rather than do this, we examine an alternative diagnostic: plots of the power spectrum. As briefly discussed in Section S1 of the Supporting Information, the use of the periodogram or even of simple direct spectrum estimates is strongly discouraged, and we apply a multitaper spectrum estimator (Thomson, 1982; Rahim et al., 2014) (with $NW=4$ and $K=7$ tapers) to the residuals instead. The result is shown in Figure 2.

Figure 2 provides an interesting perspective on the residuals. The autocorrelations shown in Figure 1 indicated that there was both a long-term trend present (the positive low-lag correlations) and something not being modeled fully (through the consistently negative correlations observed), but the spectrum demonstrates that the residuals actually have a large amount of power (variance) between periods 60 and 365 days. This range of periods corresponds to the timescales which we hoped to have accounted for by the smooth function of time. Clearly, the confounding associated with these frequencies was not fully corrected by the cubic regression splines used.

---

1A quasi-Poisson overdispersed link was also examined, with the overdispersion parameter being estimated to be 1.35. The risk estimates were identical, with slightly larger confidence intervals in the quasi-Poisson case. For demonstration purposes, the Poisson family link suffices.
Figure 1. Autocorrelation estimate [dashed, circles] for the residuals left after fitting Model 2 to 14 years of data from Chicago, using lag-0 trimmed-mean PM$_{10}$ linearly interpolated data as pollutant and a Poisson family for the link function. Also shown are results mentioned in Section 5.2, Model 10 [solid, diamonds]. The gray shaded region is a 95% confidence interval for white noise.

Figure 2. Multitaper power spectrum estimate for the residuals left after fitting Models 2 (gray, dashed) and 10 (black, solid) to 14 years of data from Chicago, Il., 1987–2000, using lag-0 trimmed-mean linearly interpolated PM$_{10}$ data as pollutant and a Poisson family for the link function. Multitaper estimate computed with $NW = 4$ and $K = 7$. The vertical dotted line shows the cutoff corresponding to 60 days, and the portion from period 14 to 2 days has been truncated for clarity.

An alternative viewpoint to the power observed to the left of the cutoff (or band-edge) at 6 cycles/year in Figure 2 would be to examine only that component of the residuals in the time domain. This could be done by use of a filter, with details provided in Sections S2 and S3 of the Supporting Information. The results provide considerable intuition: Model 2 as presented earlier leaves several obviously seasonal and yearly sinusoidal patterns in the residuals which correspond to the block of power from approximately 3 to 6 cycles/year in Figure 2. Removal of this residual long timescale variation is the core objective of the rest of this paper.

3. SOLVING SEMI-PARAMETRIC GAMS

3.1. Smoothers as filters

When using time-based smoothers in GAMs, the solution becomes a function of the smoother matrix $S$:

$$S = X(X'X)^{-1}X'$$

with $X$, the matrix whose columns are basis vectors for the subspace $S$ represented by the smoother $S$, and with superscript $'$ used to denote matrix and vector transposition. Thus, $Sx$ is a projection of $x$ onto the subspace $S$. Considering the smoothers as frequency-domain filters, appropriately chosen basis vectors yield a subspace with significant concentration in low frequencies below a chosen bandwidth $W$ (Section 4.1), the smoother known as a low-pass filter. We would like $Sx$ to be an ideal low-pass filtered copy of $x$ in the sense that it is an exact copy of the component of $x$ corresponding to frequencies less than $W$, and make this explicit in our notation by letting $x_{LP}$ denote the
ideal low-pass filtered copy of \( x \). The actual precision of the representation is dependent upon the choice of basis set for the subspace, as different representations have different concentrations, different bandwidths, and different pass-band performances. The smoother matrix \( S \) is idempotent, as is \( T = I - S \) that projects a given vector \( x \) onto the subspace orthogonal to \( S \). In filtering terms, \( T \) is a high-pass filter. Similar to the aforementioned equation, we let \( x_{\text{HP}} = Tx \) denote the ideal high-pass filtered copy of \( x \).

Note that it is impossible to create a finite-time finite-bandwidth filter. Using Slepian basis vectors for a low-frequency subspace is an optimal representation (Slepian, 1978), but is still not ideal – even in this case there will be some “bleed-through” to unwanted frequencies. Thus, the ideal filtered copies \( x_{\text{LP}} \) and \( x_{\text{HP}} \) are not attainable, and we represent this in our notation by writing \( x_{\text{LP}} \) and \( x_{\text{HP}} \) for the low-pass and high-pass filtered versions of \( x \) resulting from any finite-time projection of \( x \).

### 3.2. Solving generalized additive models with gam

For simplicity, we consider a simplified GAM rather than the model of Eqn. 1. By removing the temperature and Day-of-Week (DOW) terms from the model, while making the results uninterpretable in the context of the standard literature, we are able to simplify the derivation sufficiently for discussion. This simplified model will allow us to focus on the timescale effects clearly, and at the conclusion of this work, we will reintroduce the more complete baseline model and demonstrate similar behavior.

Using fixed \( df \) in the smooth function of time constrains the model to be a semi-parametric GAM, and can be written as a GLM. The theoretical form of the simplified GAM, we will use

\[
\log(\mu) = \begin{bmatrix} \beta_0 & \beta_1 & \beta_2 & \cdots & \beta_k \end{bmatrix} \begin{bmatrix} 1 & x^1 & x^2 & \cdots & x^k \end{bmatrix}^T
\]

(4)

where \( \mu \) and \( x \) are the vector versions of the time series \( \mu_t \) and \( x_t \), and the \( x^j \) are the basis vectors for the subspace \( S \) represented by the smoother \( S \) of Eqn. 3. In this case, the smooth function of time is a linear combination of the basis vectors \( x^j \), and will act as a low-pass filter.

The fixed-\( df \) basis set puts the estimation of \( \hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k) \) into the framework of fully linear models, so the portion of any model similar to the aforementioned equation will be solved using least squares. This will be true for any projection smoother such as natural cubic regression splines or alternatives. The findings of Dominici et al. (2002), Ramsay et al. (2003), and Dominici et al. (2004) have led to a common standardization in the use of projection smoothers, so the majority of air pollution and mortality studies will be linear in both the primary predictor of interest and the smooth function of time included to account for unmeasured confounding.

In Eqn. 4, we have a model matrix \( X = [X_1, X_2] \) that has two obvious parametric components: the mean estimate column \( I \) and pollutant predictor \( x \) forming \( X_1 \); and \( k \) smoothing columns forming \( X_2 \) (for our purposes, these are the basis vectors for a subspace, not necessarily band-limited).

When estimating solutions to GAM, typically a Newton–Raphson updating step is integral to the algorithm used (Hastie and Tibshirani, 1990, p.152–155). This is referred to as the inner loop and, for the purposes of this paper, is the area of focus. In this inner loop step, our solution devolves to that of a simple linear model, and we obtain

\[
\hat{\beta} = (X_1(I - S_2)X_1)^{-1}X_1(I - S_2)Y
\]

(5)

where \( Y \) is the vector version of the iteratively updated response, initialized as \( \mu \). Note that the \( S_2 \) matrix is computed via Eqn. 3, with \( X \) replaced with \( X_2 \). Further details of the derivation are contained in Section S4 of the Supporting Information.

### 3.3. Components of bias in the inner loop estimator

As was noted in Section 3.1, the smoother matrix \( S \) is symmetric and idempotent. For this section, we also make the assumption that the basis vectors \( x^j \) are orthonormal. While convenient for our derivation here, this assumption is irrelevant in practice because of the construction of \( S \).

Following the notation in Section 3.1 we let \( I_{\text{HP}} \) and \( x_{\text{HP}} \) denote the ideal high-pass filtered copies of \( I \) and \( x \), respectively and set \( X_{1,\text{HP}} = [I_{\text{HP}}, x_{\text{HP}}] \). In addition, \( I_{\text{HP}} \) and \( x_{\text{HP}} \) shall denote the estimates of \( I_{\text{HP}} \) and \( x_{\text{HP}} \) arising from a realizable finite time filter with \( X_{1,\text{HP}} = [I_{\text{HP}}, x_{\text{HP}}] \). Similarly, \( Y_{\text{HP}} \) shall denote the estimate of \( Y_{\text{HP}} \), the ideal high-pass filtered copy of \( Y \). Using this notation, and noting that in Eqn. 5 we have an application of the smoother \( I - S_2 \) (a high-pass filter) to \( X_1 \), which will result in \( X_{1,\text{HP}} \), and using the orthonormal basis assumption, we have

\[
\hat{\beta} = \left( X_{1,\text{HP}}' X_{1,\text{HP}} \right)^{-1} X_{1,\text{HP}}' Y_{\text{HP}}
\]

(6)

The inverse term in the aforementioned equation is that of a \( 2 \times 2 \) matrix, consisting of the product of \( X_{1,\text{HP}} \) and itself. If \( S \) were to capture the constant vector \( I \) perfectly, the solution path would be somewhat different, with the constant-capturing being aggregated with the smoother, reducing \( X_{1,\text{HP}} \) to just the column vector \( x_{\text{HP}} \) in Eqn. 6. However, with a smoother matrix failing to capture \( I \) perfectly, we write

\[
X_{1,\text{HP}}' X_{1,\text{HP}} = \begin{bmatrix} M_1 & M_2 \\ M_2 & ||x_{\text{HP}}||^2 \end{bmatrix}
\]

(7)

for constants \( M_1 = ||(I - S_2)I||^2 \) and \( M_2 = (I - S_2)I'y_\text{HP} \). As \( \hat{\beta} = [\hat{\beta}_0, \hat{\beta}_1]' \), we have \( \hat{\beta}_1 \), the inner-loop estimated coefficient for the pollutant \( x \) given by
\[ \hat{\beta}_1 = \frac{M_1 \hat{S}_{\text{HP}} \hat{Y}_{\text{HP}} - M_2 \left( \hat{\mathbf{I}}_{\text{HP}} \hat{\mathbf{Y}}_{\text{HP}} \right)}{M_1 ||\hat{S}_{\text{HP}}||^2 - (M_2)^2} \]  

(8)

Note that if our smoother were to capture constants perfectly (as discussed further in Section 4.3), then \( \mathbf{I}_{\text{HP}} = 0 \), and the \( 1 \times 1 \) matrix would give \( \hat{\beta}_1 = \hat{S}_{\text{HP}} \hat{Y}_{\text{HP}} / ||\hat{S}_{\text{HP}}||^2 \) (which should be apparent as the goal of the simplified model). As our smoother cannot be perfectly band-limited (Slepian, 1978) even with the constant capturing property, we would still not obtain our ideal result of \( \hat{\beta}_1 = \hat{S}_{\text{HP}} \hat{Y}_{\text{HP}} / ||\hat{S}_{\text{HP}}||^2 \).

Instead, rewriting \( \hat{\beta}_1 \) as

\[ \hat{\beta}_1 = \frac{\hat{S}_{\text{HP}} \hat{Y}_{\text{HP}}}{||\hat{S}_{\text{HP}}||^2} - \frac{M_1 ||\hat{S}_{\text{HP}}||^2}{M_1 ||\hat{S}_{\text{HP}}||^2 - (M_2)^2} - \frac{M_2 \left( \hat{\mathbf{I}}_{\text{HP}} \cdot \hat{\mathbf{Y}}_{\text{HP}} \right)}{M_1 ||\hat{S}_{\text{HP}}||^2 - (M_2)^2} = \alpha \frac{\hat{S}_{\text{HP}} \hat{Y}_{\text{HP}}}{||\hat{S}_{\text{HP}}||^2} - \gamma \]  

(9)

We see that the three components for the error in the inner loop estimator can be identified: the multiplicative error stemming from \( \alpha = M_1 ||\hat{S}_{\text{HP}}||^2 / \left( M_1 ||\hat{S}_{\text{HP}}||^2 - (M_2)^2 \right) \), the additive error \( \gamma = M_2 \left( \hat{\mathbf{I}}_{\text{HP}} \cdot \hat{\mathbf{Y}}_{\text{HP}} \right) / \left( M_1 ||\hat{S}_{\text{HP}}||^2 - (M_2)^2 \right) \) and the error in the estimation of \( \mathbf{S}_{\text{HP}} \) and \( \mathbf{Y}_{\text{HP}} \) due to the non-ideal nature of any chosen filter. The last of these will vary with the choice of smoother (and its bandwidth, see Section 4.1) and depend largely on the performance and properties of the same. In general, spline-based filters will have poor out-of-band performance (because they leave too much variance in low frequencies) and will be simultaneously missing some of the required in-band variance. Unfortunately, we cannot simulate or compute or implement ideal filters, so the best we can do is visually compare certain aspects of chosen filters with those of an ideal (in a concentration-of-energy sense). We will examine these points further in the following sections.

4. BIAS CORRECTION

4.1. Time-bandwidth

In many estimates of short-term risk due to air pollution, smooth functions of time are included as fixed-\( df \) cubic regression splines. Cubic splines are an obvious choice given the typical formulation of the problem, because they include a penalization term that penalizes “wiggliness” (Wood, 2006). The restriction to fixed-\( df \) splines rather than penalized splines (which have adaptively chosen degrees-of-freedom) has been done to avoid the issue of concavity (Ramsay et al., 2003; Dominici et al., 2004). The default choice for the field has generally settled on 6–8 \( df \) per year for fixed-\( df \) cubic regression splines [see, e.g., (Zanobetti and Schwartz, 2009; Samoli et al., 2011; Bell et al., 2014) (6 \( df \) per year), (Shin et al., 2012) (7 \( df \) per year), and (Peng et al., 2008; Chang et al., 2011) (8 \( df \) per year) as recent examples from the literature].

Appearing simultaneously in the literature with the choice of fixed 6–8 \( df \) per year cubic regression splines is the interpretation that such smooth functions of time account for temporal variation with periods 2 months and longer [e.g., (Dominici et al., 2000; Peng and Dominici, 2008)]. The logic is that 6 \( df \) (or knots) per year corresponds to approximately 2 months between knots, and thus, these splines should capture such variation, or most of it, leaving residuals that are of shorter temporal structure. However, this interpretation, while not completely incorrect, is flawed if taken at face value, as such splines are not designed to specifically control for such variation. It is approximately correct but with little direct control or understanding of the approximation error, which leads to uncertainty in how the risk estimates should be interpreted.

To examine these splines further, consider Figure 3, which contains several magnitude transfer functions. These are functions of frequency that present the ratio of two power spectra: the power spectrum of the output from a filter to the power spectrum of the input to a filter. The interpretation is straightforward: regions with magnitude equal to 1 perfectly transfer from the input to the output, regions with magnitudes significantly lower than 1 are attenuated (suppressed), and regions with magnitudes significantly larger than 1 are amplified. We desire an ideal transfer function (the dashed blue line in Figure 3) that perfectly transfers periods longer than 2 months (frequencies less than about 6 cycles/year), and perfectly attenuates all other periods, so that the smooth function of time will capture the long timescale variation, leaving the short timescale variation untouched. The black curve in Figure 3 is the theoretical transfer function for a 10-year 6 \( df \) per year natural cubic regression spline basis, while the grey curve is the average (over \( M = 1000 \) simulations) empirical transfer function for white noise data of the same length, 10 years (\( N = 3650 \) samples). The reader should carefully note the logarithmic scale on the \( y \)-axis of this figure.

Given a basis set for a spline defined over time sampled in days, with knots spaced evenly every \( n \) days, the linear filter (the smoothing matrix \( \mathbf{S} \) in Eqn. 3) has two properties that are of interest to us. The first property is the placement of the first zero, which in Figure 3 is the first sharply plunging point for the black curve at frequency 6 cycles/year. This property has correctly been interpreted by previous analysts, and using an \( n \)-day spacing corresponds to a first zero at a period of \( n \) days (e.g., 60-day spacing of knots equals first zero at 60 days period).

The second property, however, has been neglected: the behavior of the filter from 0 to the first sharply plunging point for the black curve at frequency 6 cycles/year. This property has correctly been interpreted by previous analysts, although the coefficients obtained from previous analyses were not incorrect, any interpretation attached to them which discussed the “short-term risk” and implied a temporal cutoff of 2 months should be re-thought. The decay of the magnitude-response curve in Figure 3 (in grey) shows that such filters (with 6–8 \( df \) per year) are perfectly capable of capturing features with temporal structure of 120 days and longer, but that much of the structure between 60 and 120 days will not be captured by the smooth function of time and thus will be left in the residual after accounting for it.
Increasing $n$ in the aforementioned equation would push higher (in frequency) the first zero of the cubic regression spline transfer function, allowing the poor magnitude performance of the latter half of the pass-band to be applied to data that should be attenuated by the filter, and the better performance of the prior half of the pass-band to apply to temporal variation at or around 3–4 months, which includes the seasons. This has been investigated with 12 $df$/year natural spline smoothers in the literature [see, e.g., (Farhat et al., 2013; Peng et al., 2013)]. However, this approach is not completely satisfactory because of the extremely slow decay of the magnitude transfer function of natural cubic regression splines and does not protect against confounding by seasonal variation.

We close this section with a brief note about bandwidth. For a competent filter, the bandwidth is a parameter of the filter that describes the position of the band-edge, which typically occurs at or around the first zero. Thus, for a fixed-$df$ cubic regression spline basis with knots spaced every $n$ days, the bandwidth is equal to $\approx 365/n$ cycles/year. This parameter is of interest because choosing its value a priori will allow us to then design an appropriate filter to obtain it, regardless of the basis choice. This is discussed further in the next section.

4.2. Alternative smooth function of time

We now introduce an alternative family of finite-time basis vectors for the frequency-limited subspace described by the traditional cubic regression spline smoother. This family has improved pass-band performance, band-edge performance, and other additional improvements over the traditional choice.

We mentioned previously that cubic regression splines arise naturally from an optimization problem with a penalty on “wiggliness”. Their primary advantage is simplicity: they can be computed quickly and efficiently, and their knot-based creation technique is highly flexible. From our point of view, their primary disadvantages are poor pass-band performance, weak out-of-band attenuation properties, and non-obvious links between bandwidth and the dimension of the basis set, which is the total $df$ of the smoothing splines. Thus, when seeking an alternative to cubic regression splines for use as time-based smoothers, we would like to address the disadvantages. Fortunately, all three disadvantages can be resolved although at the cost of some computational complexity.

In a series of papers beginning with Slepian and Pollak (1961) and ending with Slepian (1978), the authors developed the full theory of the prolate spheroidal wave functions and sequences (now referred to as Slepian). These functions arise naturally out of a particular set of spherical equations and have the unique property that among all finite-length sequences, they are the most concentrated in frequency. That is, for a given bandwidth $W$, a set of Slepian sequences have the most concentrated band-pass region $[0, W]$ of any possible basis set. Thus, we obtain a finite-time linear projection filter with flat pass-band, good band-edge performance, extremely strong out-of-band attenuation, and an obvious (formulaic) link between the bandwidth and the dimension of the basis. Essentially, all of the disadvantages of the cubic regression splines are solved at once. In addition, the Slepian sequences have the property that they are full span; by this, we mean that each basis vector is defined on $[1, N]$. This contrasts with the cubic regression spline basis vectors, which are zero except on their individual locally defined spans. This has implications for smoothing of data with missing values. The details of the generation of the Slepian sequences are beyond the scope of this paper, so we point the reader to Slepian (1978) and Thomson (1982) for explicit algorithms and two $R$ packages, (Rahim et al., 2014; Burr, 2014), which include routines for computation.

To uniquely identify a set of basis vectors, three parameters are required for a Slepian basis. The Slepian parameters are the series length ($N$), the bandwidth ($W$), and the desired number of basis vectors ($K$); for cubic regression splines, the parameters are the length ($N$) and the $df$, or number of knots, or equally, the number of basis vectors ($K$). However, the three Slepian parameters are related when the optimal filtering performance is desired, and it can be shown that in this case, we must have $K \approx 2NW$. Note that the units of $N$ and $W$ must be
comparable, so if \( N \) is the number of days, then \( W \) must be in cycles-per-day. From a practical viewpoint, given \( N \) (fixed by one’s data) and a desired temporal cutoff of \( n = 365/W \) days, the number of basis vectors \( K \) is then determined as \( 2 \cdot N \cdot W \).

Figure 4 shows several magnitude transfer functions for two choices of \( K \) with \( N = 3650: K = 60 \) (with bandwidth \( W = 3/365 \) cycles/day and \( \approx 122 \) days period) and \( K = 120 \) (with bandwidth \( W = 6/365 \) cycles/day and \( \approx 61 \) days period), one per \( K \) from both cubic regression splines and Slepian sequences. Note the straight band-edge (drop-off) for both grey curves in Figure 4 and the close resemblance to the ideal rectangular shape. Additionally, both Slepian sequence transfer functions have extreme out-of-band attenuation (very low values after the initial drop-off), as compared with the cyclic rise-and-fall nature of the cubic regression spline transfer functions. All curves in Figure 4 are theoretical, not empirical, and again note the logarithmic scale for the \( y \)-axis.

From a purely graphical standpoint, it appears clear that the Slepian sequences have significantly improved properties when compared with the cubic regression spline option, especially when the two use the same \( df \). However, as discussed in Section 3, our real goal with this replacement is to minimize the impacts of model bias on the estimation of \( \beta_1 \). Of the three sources of error identified in Section 3.3, the smoother change suggested here directly resolves only the third source of error: the capturing of the true \( x_{HP} \) and \( Y_{HP} \). As the Slepian filter is better than the cubic spline filter in every important respect, the \( Y_{HP} \) obtained with Slepian sequences will be closer to the true \( Y_{HP} \). Figure S1 from the Supporting Information shows that the difference between the two smoothers will be a smooth, slowly fluctuating function of time. In the next section, we will attempt to minimize \( M_1 \) and \( M_2 \) so as to minimize the two error sources \( \alpha \) and \( \gamma \) from Section 3.3.

### 4.3. Minimizing multiplicative (\( \alpha \)) and additive (\( \gamma \)) error

In Section 3.3, we labeled two components of error from the inner-loop estimation as \( \alpha \) and \( \gamma \) and derived both as functions of \( M_1 \) and \( M_2 \), \( x_{HP} \) and \( Y_{HP} \). In the previous section, we suggested an alternative time-based smoother that will improve our estimates of \( x_{HP} \) (and thus of \( Y_{HP} \)). In this section, we propose a modification to the Slepian sequences that will allow us to minimize \( M_1 \) and \( M_2 \), which are both dependent upon the chosen time-based smoother, through its performance capturing constants and performing the filtering operation. Both \( M_1 \) and \( M_2 \) are functions of how well the chosen time-based smoother can capture a constant. Most linear filters do only passably well at capturing constants, including the Slepian sequences. Unless explicitly designed for, projecting a constant vector onto the subspace of a linear filter will result in Gibb’s ripples (Hewitt and Hewitt, 1979) around the constant level. A visual examination of the ripples from both cubic regression splines and Slepian is included in Section S6 of the Supporting Information. Thus, any relation (such as \( M_1 \) and \( M_2 \)) that contains an expansion of \( I \) on a basis set from a linear smoother will contain a slowly oscillating term, unless explicitly designed around.

We propose here to modify the Slepian sequence basis, while retaining the majority of their desirable properties, in order to allow them to either pass or capture constants to within numerical precision. While it is theoretically possible to modify fixed-\( df \) cubic regression splines to also pass or to capture constants, this happens at the expense of both computational efficiency (one of their few clear advantages) and transfer function effectiveness (a property that is already poor).

As discussed by Thomson DJ (2001), beginning with a basis set of \( K \) Slepian sequences, a constraint is added so that the projection of an arbitrary time series onto that basis set will be decomposed into two parts: the “random” part (what we call \( Y_{HP} \)) and the “deterministic” part (for our purposes, the constant estimator). The solution lies in restructuring the basis vectors into an alternative linear combination so that the first basis vector becomes a constant vector, and the others are orthogonalized around the first. This retains the highly concentrated subspace property and the majority of the pass-band and band-edge properties demonstrated in the previous section, while simultaneously making the new projection smoother matrix perfectly reconstruct constants. This, in turn, makes \( M_1 = 0 \) and \( M_2 = 0 \). For clarity, we will differentiate between three possible choices for \( S_2 \), the projection smoother matrix: \( S \)-NS for cubic regression splines, \( S \)-SLP for unmodified Slepian sequence smoothers, and \( S \)-SLP2 for the modified variant of the Slepian sequence smoother.
An implication to the construction of S-SLP2 is that as one of the basis vectors for the smoother is now a constant vector, there is an identifiability problem between the smoother and the intercept term in the GAM. This is a similar problem to that caused when the basis vectors of the smoother are not zero-mean (see, e.g., (Wood, 2006, Ch. 4)). If left untouched, the algorithms used to solve GAMs will null out the constant column, removing its effect entirely. We can acknowledge that there is an identifiability issue, and instead of attributing the mean to the intercept, remove the intercept entirely and allow the smooth function of time to capture this effect. This latter has two benefits: it ensures that we know exactly how and why the constant intercept has been allocated and it retains the constant-capturing effect for use in creating $\hat{X}_\text{HP}$. Recall that the smooth function of time is applied to both the response and all parametric predictors, and whereas the response has the intercept term available for capturing its constant effect, the predictors do not. This is useful, as we will discuss in the final example of this paper.

If the intercept term is removed from the model, and the smooth function of time is allowed to capture all mean effect in the response, the model simplifies. Equation 6 still holds, but with $\hat{X}_\text{HP}$ replaced simply by $\hat{X}_\text{HP}$: no constant column vector. This gives $\hat{\beta}_1 = \frac{\hat{X}_\text{HP} Y_{\text{HP}}}{||\hat{X}_\text{HP}||^2}$, the intuitively appealing result. This approach completely eliminates $M_1$ and $M_2$ from consideration, and in computing $\hat{\beta}_1$ leaves as the only source of error the efficacy of the linear filter in capturing $\hat{X}_\text{HP}$. Thus, using S-SLP2 solves all of the problems discussed in Sections 3.3 and 4.2: the pass-band and band-edge issues with the natural cubic regression spline smoother, the $\alpha$ and $\gamma$ error that arises through estimation of the mean function, and the bandwidth interpretation problem.

Alternatively, a slightly reduced-$df$ variant of S-SLP2 can be used, leaving the intercept term intact. In this case, the first column of the modified basis set is dropped, removing any identifiability issue, and changing the smoother matrix from reproducing constants to ignoring them. We refer to this variant as S-SLP3. Explicitly, this changes from S-SLP2 $\cdot 1 = 1$ to S-SLP3 $\cdot 1 = 0$. This method also eliminates $M_2$ and sets $M_1 = 1$, leading to the same intuitively appealing result as aforementioned earlier, while retaining the more traditional intercept option. In both cases, the total number of $df$ is the same, and the performance of the two smoothers are the same except for their constant-capturing difference. The one advantage of S-SLP2 over S-SLP3 is the potential for capturing constant offsets in the parametric predictors. Notably, many air pollutant series are approximately zero-mean by design, but not perfectly so, especially near the endpoints of the series. Allowing the smooth function of time to capture their mean offset allows for better performance of the filter near the endpoints, a feature that may compensate for the reduced filtering power in the middle of the time span.

A package called 1p (Burr, 2014) for the R programming language (R Development Core Team, 2012) has been created that constructs basis objects for the different variations of Slepian sequences mentioned in this paper. This package includes routines to support both gam (Hastie, 2013) and mgcv (Wood, 2011) as these are the two most commonly used GAM solvers. The package is available on the Comprehensive R Archive Network.

4.4. Summary

When considering smoothers for short-term risk models, our process was threefold. First, we demonstrated that the target bandwidth is not met through the use of 6 $df$/year smoothers (Section 4.1), regardless of family. This can be corrected through the use of increased $df$/year, with 12 $df$/year corresponding to a timescale of 2 months. Second, we introduced a new family of smoothers, the Slepian sequences (S-SLP-), that corrected the poor band-edge and pass-band properties of the spline families, but went no further (Section 4.2). This second stage corrected only one underlying bias issue, and had little-to-no effect on $\alpha$ and $\gamma$. Finally, in the third stage (Section 4.3), we modified the standard Slepian family in order to remove the remaining bias, resulting in our S-SLP2 family of smoothers.

The improvements gained from the second and third stages of our process are significant. While it may be tempting to simply use S-NS smoothers with increased $df$, we remind the reader that most current publications observe little-to-no difference between 8 and 12 $df$/year when using S-NS, indicating that the two protect against confounding equally poor. Simply increasing the bandwidth of the smoother is not sufficient to protect against seasonal variation, but using a more efficacious smoother is. The second stage of our process is critical, as it eliminates an underlying and underappreciated problem with spline smoothers. This will be demonstrated further in the simulations in Section 5.1. In addition, the Slepian’s have the advantage that the interpretation for bandwidth is inherent to their construction, so that choosing $a$ priori a cutoff for a research definition of “acute risk” equates to choosing a precise $W$, which in turn equates to a $df$ ($K$). This relationship is not clear when using S-NS, as is evidenced by the interpretations attached to S-NS-6 in current research.

5. RESULTS

In this section, we will apply and contrast our three new smoothers, S-SLP, S-SLP2, and S-SLP3 with more traditional S-NS cubic regression spline smoothers. The discussion is in two parts. In the first part, we will demonstrate the necessity of adjusting the bandwidth $W$, as discussed in Section 4.1, and we will contrast the results of the different smoothers on a number of synthetic data sets (Section 5.1). In the second part (Section 5.2), we will apply the various smoothers to a real-world data set from Chicago, Il, using the NMMAPS database (Peng and Welty, 2004). The only modification made to the data from this database was to linearly interpolate the small number of missing points in the PM$_{10}$ series.

5.1. Simulation study: traditional versus proposed smoothers

In this section, we endeavor to demonstrate the error terms of Section 3.3 through the use of simulated data. In doing this, we will compare and contrast the various smoothers we have introduced in this paper with more traditional choices, such as natural cubic regression splines with 6 and 12 $df$/year. Eight smoothers are used: S-NS-6 and S-NS-12, S-SLP-6 and S-SLP-12, S-SLP2-6 and S-SLP2-12, and S-SLP3-6 and S-SLP3-12, as denoted in the previous section. Further details of the simulations are given in Section S5 of the Supporting Information.
We use four simulations, with each in turn adding complexity to a common base. To ensure that we are able to examine the true high-pass components of the predictor and response, we simulate background noise for the realizations by using a family of sinusoids, evenly spaced in frequency from 0 cycles/day to 0.5 cycles/day. This covers the band applicable to data sampled daily, as is typical in environmental health data. For sim1 and sim2, the sinusoids are scaled by amplitudes that are randomly sampled Gaussian noise realizations, while for sim3 and sim4 the Gaussian noise realizations have variance exponentially decaying with frequency: sim3 and sim4 are more closely related to the background noise of typical mortality and pollution series, which tend to have significant power for timescales longer than 60 days.

In either case, the noise sequences have specific signals added to both the low-pass and high-pass regions. Full details of the added signals are provided in Table S2 of the Supporting Information. Note that the only difference between sim3 and sim4 is that the noise and signal portion of the Y (response) series have been scaled by a factor of 2. This accentuates any ill-fit, driving the obtained risk coefficient down used in models 1:0 and 0:W. BURR, G. TAKAHARA AND H. SHIN and smoother with bandwidth less than 6 cycles/year (corresponding to 12 cycles/year) is that the noise and signal components of the realization (X1,HP and Y1,HP), and the mean and primary regression coefficients (β0 and β1) of a GAM computed on the full realization series with the appropriate time-based smoother included. The GAMs used to share the Gaussian family due to issues with the Poisson and quasi-Poisson family that are discussed in Section S5 of the Supporting Information. The results provided are arithmetic means over the 250 realizations.

In Table 1, we see that the performance of the estimators depends heavily upon both the choice of bandwidth and the relationship between the response and predictor in the region between the interpreted bandwidth and the actual bandwidth. In the case of 6 df/year smoother, if there is a strong correlation between mortality and air pollution in the period from 2 to 6 months (which there typically is, both series being strongly seasonal in nature; in the simulations, sim3 and sim4 both qualify), this will influence the risk estimate, which then cannot be interpreted as being “short-term” or local in timescale. In cases where the two series are essentially uncorrelated in this region, or where the correlation is the same as the high-frequency range, the impact will be reduced. Our conclusion from this small demonstration is that bandwidth matters, and if we wish to be able to interpret our results as being short-term, we must use appropriate smoothers that filter with bandwidth that corresponds to our interpretations. At a minimum, 6 df/year (approximately 3 cycles/year bandwidth) smoothers of any type are not sufficient to protect against long timescale or seasonal confounding, regardless of any other properties. This explains the poor performance of the Slepian family smoothers with bandwidth corresponding to 6 df/year.

Simulation 1 is the simplest case considered. The coherent structure for timescales longer than the cutoff of 60 days consists of two periodicities at periods 183 and 75 days. Eight smoothers are capable of attenuating the 183-day signal, so differences will be due to the performance on the 75-day signal. We see the bandwidth result immediately: the first four smoothers (ending in -6) all dramatically underestimate β1 = 1 because they fail to protect against the negative correlation at 75 days period, while the last four results protect against this, giving results that are much closer to the true β1.

Simulation 2 extends Simulation 1 by adding three additional signals, two of which are in the region from 60 to 120 days period where natural cubic regression spline smoothers have poor performance (refer again to Figure 3). As we can see from the results, the signals added at periods 68 and 105 days have a demonstrable effect with smoother S-NS-12, S-SLP2-12, S-SLP3-12. Again the impact of the bandwidth choice is clear: all -6 smoothers have been driven to near-zero by the strongly negative correlation of the 2–6 months region.

Simulation 3 uses a similar signal structure as Simulation 2, but changes the background noise to have an exponential decay. As the noise is uncorrelated between the predictor and response, this has little impact on the overall results, and the coefficients obtained for the -12 smoothers are comparable with those of Simulation 2. The results for the -6 smoothers do change, but this is due to a large amount of the decaying noise being included in the final regression computation, balancing out the strongly negative correlation of the signals with a near-zero correlation (and a significant variance contribution). The -6 smoothers are all badly biased.

Finally, Simulation 4 uses the same signal structure and exponentially decaying background noise of Simulation 3, but scales by a factor of both the two background noise (via the variance) and signals (via amplitude). The results are immediately evident for the -6 smoothers and S-NS-12, driving the bandwidth results from the -6 smoothers fully negative, and dropping the β1 for S-NS-12 down to 0.806, almost 20% below the true estimate β1 = 1.0.

The total df used in models S-SLP2-12 and S-SLP3-12 is the same, with the only difference being the attribution of the intercept term. The main advantage of using S-SLP2-12 over S-SLP3-12 is that it allows the model to accurately estimate and remove the mean effect of the x term. This has implications for the performance of the chosen time-based smoother when it is applied to X1 and should be an advantage when analyzing real data.

The reader should note that sim3 and sim4 most closely resemble real-world environmental epidemiology data, with exponentially decaying background noise and significant structure in the temporal band corresponding to the seasons. This note also implies that depending on the geographic variations, significant differences in seasonal contamination will occur, so that estimates of risk from (say) Chicago will
Table 1. Simulations 1–4, using S-NS-6 through S-SLP3-12

| Smoother | $\alpha - 1$ | $\gamma$ | $\lambda$ Result | $\hat{\beta}_0$ | $\hat{\beta}_1$ |
|----------|--------------|----------|-------------------|-----------------|-----------------|
| S-NS-6   | 3.550e-03    | -6.301e-03 | 2.806e-01         | -2.957e+00      | 2.841e-01       |
| S-NS-12  | 1.059e-03    | -4.190e-03 | 9.373e-01         | -4.508e+00      | 9.322e-01       |
| S-SLP-6  | 2.424e-03    | -1.529e-03 | 1.418e-02         | -4.796e-01      | 4.232e-02       |
| S-SLP-12 | 1.499e-02    | 2.910e-03  | 9.977e-01         | 9.925e-01       | 9.869e-01       |
| S-SLP3-6 | 8.005e-05    | 1.810e-05  | 4.182e-02         | 1.007e+00       | 4.105e-02       |
| S-SLP3-12| 1.945e-04    | 4.398e-05  | 9.858e-01         | 9.810e-01       | 9.709e-01       |
| S-SLP2-6 | 1.864e-02    | -2.013e-02 | 4.453e-02         | -5.969e+00      | 6.416e-02       |
| S-SLP2-12| 1.191e-01    | 1.009e+00  | 1.191e-01         | 1.168e-01       | 1.168e-01       |
| S-SLP3-6 | 5.056e-03    | -1.365e-02 | 8.554e-01         | -1.089e+01      | 8.373e-01       |
| S-SLP3-12| 1.195e-02    | 2.225e-03  | 9.883e-01         | 6.078e+00       | 9.823e-01       |
| S-SLP2-6 | 2.362e-03    | -1.370e-03 | 1.191e-01         | -6.880e-01      | 1.182e-01       |
| S-SLP2-12| 1.194e-04    | 2.564e-05  | 9.882e-01         | 9.830e-01       | 9.722e-01       |
| S-SLP3-6 | 1.502e-04    | 1.052e-05  | 9.977e-01         | 9.925e-01       | 9.869e-01       |
| S-SLP3-12| 4.263e-04    | 5.680e-05  | 9.882e-01         | 9.830e-01       | 9.722e-01       |
| S-SLP2-6 | 2.270e-03    | -2.969e-03 | 2.129e-01         | -1.253e+00      | -2.118e-01      |
| S-SLP2-12| 1.855e-04    | 2.300e-05  | 9.902e-01         | 5.086e+00       | 9.785e-01       |
| S-SLP3-6 | 4.820e-03    | -2.798e-02 | 7.906e-01         | -2.036e+01      | 8.058e-01       |
| S-SLP3-12| 1.940e-04    | 5.105e-05  | 9.902e-01         | 9.869e-01       | 9.701e-01       |
| S-SLP2-6 | 1.854e-04    | 2.300e-05  | 9.902e-01         | 9.869e-01       | 9.701e-01       |
| S-SLP2-12| 4.063e-04    | 5.105e-05  | 9.902e-01         | 9.869e-01       | 9.701e-01       |

Estimates of $\alpha$ and $\gamma$ are irrelevant for S-SLP2- smoother so are left blank. The $\alpha - 1$ results are presented for easy comparison, as all $\alpha$ are 1.0 rounded. The $\lambda$ column gives results for a linear model estimate on the true realizations, showing that some of the difference between the results in the final two columns and the true $\beta_1$ is due to the structure of GAM solvers. The shaded cells are the comparisons (per simulation) between classic versus proposed smoothers.

be biased differently to estimates from Los Angeles or Seattle. Thankfully, the use of the S-SLP2-12 smoother compensates for this seasonal geographic variation, allowing for a standard model to be used. This note and other implications of these simulations are continued in Section S5 of the Supporting Information.

5.2. The NMMAPS data set: Chicago, 1987–2000

As mentioned earlier in the paper, we now expand our scope to consider a more complete model as might be used in an analysis of real-world air pollution and health data. Specifically, we return to Eqn. 2, and for the city of Chicago, II., use the 1987–2000 time span of data, as provided by the NMMAPS data database (Peng and Welty, 2004). We restrict attention to two pollutants (trimmed mean daily PM$_{10}$ and trimmed mean daily Ozone) and two health outcomes (daily all-cause mortality counts – Death and daily cardio-vascular mortality counts - CVD). Terms for Day-of-Week and a smooth function of temperature with 3 df are included. Finally, eight smooth functions of time are considered, as in the previous simulation discussion: S-NS-6 through S-SLP3-12, with the -6 smoothers using 6 df/year and the -12 smoothers using 12 df/year. The numerical results for the possible models considered are given in Table 2.

Examining the results in Table 2, we can clearly see differences in the risk estimates for changes in smoother (NS to SLP or SLP2) and bandwidth (from the -6 smoothers group to the -12 smoothers group) for the four pollutant/health effect combinations. In three of the four
combinations, when directly comparing S-NS-6 with S-SLP2-12, the change is towards zero (that is, the original result could be considered an overestimate), but we caution the reader not to over-generalize these results immediately. All that can be concluded from this table is that for the three combinations that decrease; the relationship between predictor and response (possibly confounded with temperature) is stronger at timescales between 2 and 4 months than it is at timescales shorter than 2 months. Recall that the primary difference between the S-NS-6 smoother and any of the Slepian S-SLP2 or S-SLP3 classes is the bias caused by unattenuated power at medium timescales. Thus, the difference observed must be due to some interaction with this power.

For purposes of demonstrating the impact of moving from a traditional $df$ choice of cubic regression spline smoothers (in particular, S-NS-6 used as an example) to Slepian mean-adjusting smoothers (using S-SLP2-12 as our example), consider Figure 5, where risk estimates with 95% confidence intervals are presented for S-NS-6 and S-SLP2-12 for all four combinations of pollutant/health outcome considered in Table 2. This comparison corrects the bandwidth issue (removing all seasonality) by moving to -$12$, and removes the error because of $\alpha$, $\gamma$, and mean-estimation by moving to the S-SLP2 family. The S-SLP2-12 smoother results shown in Figure 5 can accurately be said to estimate the risk (for chosen health outcome and pollutant) from acute timescales up to 2 months, with minimized bias due to seasonality and smoother construction. The S-NS-6 smoother results cannot be said to achieve any of these things, most especially the acute timescales point.

When comparing the performance of the various smoothers in Table 1 with the results in Table 2, it is reasonable to consider why S-SLP-6 is not similar to S-SLP2-6 and S-SLP3-5, when in the simulations the results are quite comparable for $\hat{\beta}_1$. The reason for this difference is contained in our minimization of $\alpha$ and $\gamma$ errors, and in particular, in the interaction of the smooth function with non-zero means. Correcting for these was the impetus behind the creation of the S-SLP2 and S-SLP3 smoothers. Thus, when examining the results for Ozone and all-cause mortality (Death) in Table 2, we can attribute any difference between S-SLP-6 and S-SLP-2 to the interaction of the non-zero means of the smooth function of temperature, the ozone, the log mortality series, and the interactions of the seasonal variation in each of these. S-SLP-6 is a more accurate estimator of long timescale structure and is a better representation of the underlying true risk attributed to ozone exposure at timescales up to 4 months; S-NS-6 and S-SLP-6 are both over-estimates of the same. It is merely coincidence that the risk attributed to ozone exposure at timescales up to only 2 months (the -12 smoothers) is a similarly valued quantity.
Environmetrics

Figure 5. Generalized Additive Model coefficient estimates for Chicago, Il., 1987 – 2000. Each row shows the risk estimate and 95% confidence interval for the indicated health outcome and pollutant, with the risks computed using models that include terms for Day-of-Week and a smooth function of temperature, as well as a smooth function of time, as noted to the right of the plotted intervals. The health outcomes are All-Cause (Death) and Cardio-Vascular (CVD) mortality.

It is quite difficult to generalize from these initial results, especially given that different geographic regions will respond differently to seasonal variation, and will have different relationships between predictor and response in the medium timescales affected by the time-based smoother bias we have been discussing. One advantage of our new point of view is that it gives us more tools for examining these relationships; for example, precise cutoff periods are now possible, so if an analyst were interested in estimating risk due to acute exposure out to (say) 90 days, using the bandwidth corresponding to \( \approx 8.1 \) df/year would allow such an estimate to be computed (Section 4.1 for precise details). Thus, it may be possible to refine the standard approach for choice of df so as to isolate particular effects that are of interest for study. We remind the reader at this point that the -6 smoothers by construction should only be intended to restrict attention to timescales up to 4 months, not the 2 months that is commonly attached in interpretation to S-NS-6. This is the primary reason why -12 smoothers are considered here: especially when using S-CLP-class smoothers, control of bandwidth is quite precise, and if we wish to attach an interpretation to the results that refers to a timescale of “up to 2 months” (or similar), we must use -12 smoothers, not -6 smoothers. In addition, recall the discussion surrounding Figure 4 where we showed that the use of S-NS-class smoothers will never have an immediate relation between choice of df and precise bandwidth, because of the smearing caused by poor frequency concentration in the transfer function. Use of S-CLP-class smoothers solves this problem.

We now close this section by re-examining Model 2 with our updated S-CLP-12 time-based smoother:

\[
\log(\mu_t) = \beta_0 + x_t \beta_1 + \text{DOW}_t \beta_2 + \mathcal{S}(\text{Temp}, df = 3) + \mathcal{S}(\text{Time}, df = 12/\text{yr})
\]

with the smooth function of temperature remaining a natural cubic regression spline, but with the smooth function of time being the S-CLP-12. The results from this Model 10 are included as secondary lines in Figures 1 and 2, and Figure S1 in the Supporting Information. In Figure 1, comparing the two autocorrelations, it appears that the use of Model 10 has caused the positive correlation at low lags to be reduced (indicating less seasonality), and the significance bound is exceeded at fewer lags. Insofar as autocorrelations can be interpreted (which is to say, only at a very coarse level), the improvement appears to be a success. Note that we do not expect the results to fully resemble white noise, as they are not noise.

Figures 2 and S1 (Supporting Information) are more interesting. In Figure 2, we see strong evidence from the spectrum (black curve) that our S-CLP-12 smoother has been effective at its task: removing all variation (power) at timescales longer than 60 days. The band-edge is sharp, the attenuation for out-of-band power is strong, and the placement of the band-edge is appropriate at the frequency shown as a dashed line (60 days). Figure S1 in the Supporting Information showing the time-domain residuals split into long-timescale and short-timescale orthogonal components is also interesting, with little-to-no long-timescale variation remaining in the residuals (thick black curve), as compared with the original S-NS-6 model (thick dark grey curve). The small amount that remains is impossible to fully remove, as discussed previously at the end of Section 3.1.

In summary, we have demonstrated the use of our newly defined time smoothers on real data from the NMMAPS data database and have shown a real impact from bias because of timescale contamination and smoother family. Work is ongoing to determine the impact at a regional or national level. We reiterate here that geographic variation may very well result in bias being over-estimation or under-estimation depending on the particular interaction between predictor and response.

6. DISCUSSION AND CONCLUSION

Estimation of short-term (or acute) risk due to air pollution is an important area of research and has been subject to much scrutiny in the recent past. The commonly agreed-upon model used for estimation of this risk is a GAM that includes a smooth function of time to control for unmeasured confounding. Typically, this smooth function of time is a natural cubic regression spline with low (6–8 df/year) degrees of freedom.
Through careful deconstruction of the GAM solver, we demonstrated three sources of error that may be introduced in the estimation of short-term risk, all due to the smooth function of time. We additionally showed that the use of $6 \text{ df/year}$ for a smooth function of time is not, in general, sufficient to protect estimates from seasonal variation. The use of natural cubic regression splines at higher $df/\text{year}$ (e.g., 12) will protect slightly better against seasonal variation than $6 \text{ df/year}$, but still suffers from the poor concentration properties of its family.

We proposed an alternative smoother family that reduces the three sources of error discussed in the paper. Through the use of a number of simulations, we demonstrated the improvements possible with this new family, demonstrating that a negative relationship between medium timescale predictor and response can result in an under-estimate of risk. The opposite finding is also true: a strongly positive relationship between medium timescale predictor and response can result in an over-estimate of risk. Our demonstration concluded with a discussion of a specific city: Chicago, IL, using data from 1987–2000 from the NMMAPSdata database. The results from Chicago, while not immediately generalizable, do have one concrete finding: it is almost certain that risks are being estimated with some contamination from seasonal variation. Further work will be required to examine the implications for national estimates of risk.

In conclusion, we have demonstrated a source of bias in current estimates of population health risk due to air pollution. We have proposed and demonstrated an improvement that compensates for this bias, resulting in estimates with more accurate interpretation. We have shown that the commonly used choice of $6 \text{ df/year}$ does not correspond to the commonly attached interpretation of “up to 2 months” timescale, suggesting that a more appropriate choice of $df$ may be $12 \text{ df/year}$ or higher, so long as a sufficiently concentrated smoother family is used. There is an obvious (formalistic) link between bandwidth $W$ and the basis dimension $K$ (equivalent to degrees-of-freedom) of our proposed new smoother, which allows finer control of the interpretation of the risk estimated by this smoother. Finally, our new approach provided several useful tools for the examination of timescales and risk that may provide further insights.

An R package, slp (Burr, 2014), has been written and made available via the Comprehensive R Archive Network so that our methods may be reproduced and extended. Additional information and supplementary material for this article is available online at the journal’s website, and code to reproduce all analyses, simulations, and figures are available from Burr (2015).

Acknowledgements

The authors thank Dr. Rick Burnett (Health Canada), Professor Mark Goldberg (McGill University) and Professor David J. Thomson (Queen’s University) for their helpful feedback on versions of this work. We also thank the three anonymous reviewers and Associate Editor whose comments and guidance clarified key sections and significantly improved the paper.

REFERENCES

Bell ML, Ebisu K, Leader BP, Gent JF, Lee HJ, Koutrakis P, Wang Y, Dominici F, Peng RD. 2014. Associations of PM2.5 constituents and sources with hospital admissions: analysis of four counties in Connecticut (USA) for persons $\geq 65$ years of age. *Environmental Health Perspectives* 122(2): 138–144.

Burnett RT, Smith-Doiron M, Stieb D, Cakmak S, Brook JR. 1999. Effects of particulate and gaseous air pollution on cardiorespiratory hospitalizations. *Archives of Environmental Health: An International Journal* 54(2): 130–139.

Burr WS. 2014. slp: Slepian smooth functions of time - an extension package for ‘mgcv’ and ‘gam’. v.1.0-3. http://cran.r-project.org/web/packages/slp. [Accessed 1 March 2015].

Burr WS. 2015. Code and results for “bias in estimation...” Paper. GitHub repository, https://github.com/wesleyburr/bias-in-estimation [Accessed 17 February 2015]. DOI: 10.5281/zenodo.15862.

Chang HH, Peng RD, Dominici F. 2011. Estimating the acute health effects of coarse particulate matter accounting for exposure measurement error. *Biostatistics* 12(4): 637–652.

Dockery DW, Pope CA. 1994. Acute respiratory effects of particulate air pollution. *Annual Review of Public Health* 15(1): 107–132.

Dominici F, Samet JM, Zeger SL. 2000. Combining evidence on air pollution and daily mortality from the 20 largest US cities: a hierarchical modelling strategy. *Journal of the Royal Statistical Society: Series A* 163(3): 263–302.

Dominici F, McDermott A, Zeger SL, Samet JM. 2002. On the use of generalized additive models in time-series studies of air pollution and health. *American Journal of Epidemiology* 156(3): 193–203.

Dominici F, McDermott A, Haste TJ. 2004. Improved semiparametric time series models of air pollution and mortality. *Journal of the American Statistical Association* 99(468): 938–948.

Dominici F, Peng RD, Bell ML, Pham L, McDermott A, Zeger SL, Samet JM. 2006. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *The Journal of The American Medical Association* 295(10): 1127–1134.

Farhat N, Ramsay T, Jerrett M, Krewski D. 2013. Short-term effects of ozone and PM2.5 on mortality in 12 Canadian cities. *Journal of Environmental Protection* 4(12): 18–32.

Hastie TJ, Tibshirani RJ. 1990. *Generalized Additive Models*. Chapman and Hall: London.

Hastie TJ. 2013. GAM: generalized additive models. R package version 1.09.1. http://cran.r-project.org/web/packages/gam/ [Accessed 1 March 2015].

Hewitt E, Hewitt RE. 1979. The Gibbs–Wilbraham phenomenon: an episode in Fourier analysis. *Archive for History of Exact Sciences* 21(2): 129–160.

Kelsall JE, Samet JM, Zeger SL, Xu J. 1997. Air pollution and mortality in Philadelphia, 1974–1988. *American Journal of Epidemiology* 146(9): 750–762.

Peng RD, Welty LJ. 2004. The NMMAPSData package. *R News* 4(2): 10–14.

Peng RD, Dominici F, Louis TA. 2006. Model choice in time series studies of air pollution and health. *American Journal of Epidemiology* 163(3): 179–203.

Dominici F, McDermott A, Zeger SL, Samet JM. 2006. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *The Journal of The American Medical Association* 295(10): 1127–1134.

Farhat N, Ramsay T, Jerrett M, Krewski D. 2013. Short-term effects of ozone and PM2.5 on mortality in 12 Canadian cities. *Journal of Environmental Protection* 4(12): 18–32.

Hastie TJ, Tibshirani RJ. 1990. *Generalized Additive Models*. Chapman and Hall: London.

Hastie TJ. 2013. GAM: generalized additive models. R package version 1.09.1. http://cran.r-project.org/web/packages/gam/ [Accessed 1 March 2015].

Hewitt E, Hewitt RE. 1979. The Gibbs–Wilbraham phenomenon: an episode in Fourier analysis. *Archive for History of Exact Sciences* 21(2): 129–160.

Kelsall JE, Samet JM, Zeger SL, Xu J. 1997. Air pollution and mortality in Philadelphia, 1974–1988. *American Journal of Epidemiology* 146(9): 750–762.

Peng RD, Welty LJ. 2004. The NMMAPSData package. *R News* 4(2): 10–14.

Peng RD, Dominici F, Louis TA. 2006. Model choice in time series studies of air pollution and mortality. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169(2): 179–203.

Peng RD, Chang HH, Bell ML, McDermott A, Zeger SL, Samet JM, Dominici F. 2008. Coarse particulate matter air pollution and hospital admissions for cardiovascular and respiratory diseases among Medicare patients. *Journal of the American Medical Association* 299(18): 2172–2179.

Peng RD, Dominici F. 2008. *Statistical Methods for Environmental Epidemiology with R: A Case Study in Air Pollution and Health*. Springer-Verlag: New York, NY.

Peng RD, Samoli E, Pham L, Dominici F, Touloumi G, Ramsay T, Burnett RT, Krewski D, Le Tertre A, Cohen A, Atkinson RW, Anderson HR, Katsouyanni K, Samet JM. 2013. Acute effects of ambient ozone on mortality in Europe and North America: results from the APHENA study. *Air Quality, Atmosphere and Health* 6(2): 445–453.
BIAS CORRECTION IN PUBLIC HEALTH RISK ESTIMATES

R Development Core Team. 2012. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. [ISBN] 3-900051-07-0. http://www.R-project.org.

Rahim KJ, Burr WS, Thomson DJ. 2014. Appendix A: Multi taper R Package in “Applications of Multi taper Spectral Analysis to Nonstationary Data”, PhD diss., Queen’s University, pp. 149–183. http://hdl.handle.net/1974/12584.

Ramsay TO, Burnett RT, Krewski D. 2003. The effect of concurvity in generalized additive models linking mortality to ambient particulate matter. *Epidemiology* 14(1): 18–23.

Samoli E, Nastos PT, Paliatsos AG, Katsouyanni K, Priftis KN. 2011. Acute effects of air pollution on pediatric asthma exacerbation: evidence of association and effect modification. *Environmental Research* 111(3): 418–424.

Schwartz J, Marcus A. 1990. Mortality and air pollution in London: a time series analysis. *American Journal of Epidemiology* 131(1): 185–194.

Schwartz J, Spix C, Touloumi G, Bachárová L, Barumamdzadeh T, Le Tertre A, Ponce de Leon A, Pönkä A, Rossi G, Saez M, Schouten JP. 1996. Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions. *Journal of Epidemiology and Community Health* 50(4): Supplement 1:S3–11.

Shin HH, Stieb DM, Burnett RT, Takahara G, Jessiman B. 2012. Tracking national and regional spatio-temporal mortality risk associated with NO2 concentrations in Canada: a Bayesian hierarchical 2-level model. *Risk Analysis* 32(3): 513–530.

Slepian D. 1978. Prolate spheroidal wave functions, Fourier analysis, and uncertainty – V: the discrete case. *Bell Systems Technical Journal* 57(5): 1371–1429.

Thomson DJ. 1996. Spectrum estimation and harmonic analysis. *Proceedings of the IEEE* 70(9): 1055–1096.

Thomson DJ. 2001. Inverse-constrained projection filters. In *International Symposium on Optical Science and Technology* 4478:172–183. International Society for Optics and Photonics.

Wood SN. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Methodology)* 73(1): 3–36.

Wood SN. 2006. *Generalized Additive Models: An Introduction with R*. CRC Press: Boca Raton, FL.

Zanobetti A, Schwartz J. 2009. The effect of fine and coarse particulate air pollution on mortality: a national analysis. *Environmental Health Perspectives* 117(6): 898–903.

**SUPPORTING INFORMATION**

Additional supporting information may be found in the online version of this article at the publisher’s web site.