The University of Chicago
Center for Integrating Statistical and Environmental Science
www.stat.uchicago.edu/~cises

Chicago, Illinois USA

TECHNICAL REPORT NO. 30

SEASONAL VARIATIONS IN THE SPATIAL-TEMPORAL DEPENDENCE OF TOTAL COLUMN OZONE

Michael L. Stein

November 2005
Revised: May 2006

Although the research described in this article has been funded wholly or in part by the United States Environmental Protection Agency through STAR Cooperative Agreement #R-82940201 to The University of Chicago, it has not been subjected to the Agency's required peer and policy review and therefore does not necessarily reflect the views of the Agency, and no official endorsement should be inferred.

Seasonal variations in the spatial-temporal dependence of total column ozone

Michael L. Stein
Department of Statistics
University of Chicago
5734 University Ave.
Chicago, IL 60637
stein@galton.uchicago.edu
phone: 773-702-8326
fax: 773-702-9810

Short title: Seasonal models for total column ozone

[1]Although the research described in this article has been funded wholly or in part by the United States Environmental Protection Agency through STAR cooperative agreement R-82940201-0 to the University of Chicago, it has not been subjected to the Agency's required peer and policy review and therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred.

Abstract

Summary: The Total Ozone Mapping Spectrometer (TOMS) is a satellite-based instrument that measures total column ozone on a daily basis over a fairly dense spatial grid with near global coverage. Statistical models for the spatial-temporal variations in total column ozone provide insights into ozone dynamics, are valuable for obtaining inferences on long-term trends in ozone levels, and can be used to predict ozone levels at unobserved points in space-time. However, developing such a model is complicated by the seasonally varying nature of the space-time dependence and the partial confounding of spatial and temporal variation caused by the sun-synchronous orbit of the satellite. This work considers methods for describing, modeling and estimating the seasonal patterns in the dependence structure for measurements at a single latitude. Applying one of these models to the synoptic prediction of total column ozone at a latitude shows that there is at least the potential for substantially improved predictions by exploiting the high spatial density of observations TOMS provides.

Key words: Nonstationary process, periodic correlation, spectral analysis, synoptic prediction, TOMS

1 Introduction

Satellite-based observations of atmospheric quantities provide a number of challenges for statistical modeling, especially when used at the full spatial and temporal resolution available. Among these challenges include: capturing space-time dynamics, seasonal patterns in spatial-temporal dependence, the spherical nature of the spatial domain, anomalies due to the measurement process (including missing observations), fundamental differences in dependencies at different latitudes and the overwhelming size of the datasets. This work examines a number of these issues through modeling and analysis of total column ozone levels from 1980–1993 as measured by TOMS, which provides daily values with high spatial resolution over many years. The goal here is to aim towards a comprehensive model for these data at a single latitude. The models are fitted by maximizing approximate likelihoods, which are obtained by breaking up the data into 26 two-week periods for each year, assuming the correlation structure has a period of one year, the process is stationarity within periods and is independent across periods. To approximate likelihoods within a each period, a spectral approximation is used. The resulting estimates plotted against period indicate that the data's correlation structure varies fairly smoothly with the time of year. These results suggest a valid covariance model that may capture the seasonal dependence in the correlations of total column ozone across time and longitudes.

Section 2 describes the data used in this work and reviews past analyses. The data show clear persistent differences in ozone levels across seasons and longitudes, and Section 3 describes how these influences were removed as well as how missing observations were handled. Although there is clear evidence for seasonality in the correlation structure, it does seem reasonable to treat the process as stationary within fairly short "seasons," which we take here to be two weeks long. Section 4 provides a general model for a stationary space-time process at a single latitude that includes a simple and readily interpretable way of capturing space-time asymmetries. Section 5 fits an example of this model to each of 26 two-week seasons. The model provides a good description of the data within each season. Section 6 considers the problem of synoptic prediction (i.e., prediction at a single time), which is necessary for setting initial conditions in numerical models for ozone (Lait 2000). Because this model gives the covariance function of the process at all longitudes and times, we can calculate optimal linear predictors and their mean squared errors under the fitted models. Section 6 shows how these prediction errors depend on the spatial density of the observations. Section 7 describes some initial efforts to obtain and fit a comprehensive model that captures the seasonal dependence structure and shows that with a small number of parameters, one can capture most of the seasonality found

by fitting a separate model to every two-week season. Section 8 discusses some of the computational and modeling challenges in fitting a seasonal model that addresses dependencies on time scales longer than two weeks.

2 Data

The Nimbus-7 satellite carried a TOMS instrument that measured total column ozone on a daily basis from November 1, 1978 to May 6, 1993. More recent TOMS observations are available from instruments aboard other satellites, but we will only consider this single instrument here. The Nimbus-7 satellite followed a sun-synchronous polar orbit with an orbital frequency of 13.825 orbits a day. On each orbit, a scanning mirror scanned across a track about 3000 km wide, yielding 35 measurements every 8 seconds (Krueger, et al. 1998). Thus, at least away from the poles, all observations were taken at roughly noon local time. Raw measurements were processed onto a grid of 1° latitude × 1.25° longitude for latitudes between 50°S and 50°N to obtain the Version-7 Level-3 data used here (Krueger, et al., 1998). Despite the massive effort put into producing the Version-7 data, it still has known anomalies, due, for example, to inadequate treatment of the effect of clouds (Liu, Newchurch and Kim 2003) and anomolous sea-land differences (Cuevas, et al. 2001).

Chapman, et al. (1974), Salby (1982, 1987), Hayashi (1983), Lait and Stanford (1988) and Lait (2000) all address issues arising in analyzing data taken from a polar-orbiting satellite. They largely focus on the problem of producing synoptic images from the asynoptic data the satellite-based instrument provides. Multiple observations from a single scan are not emphasized in these works because, as Salby (1987) argues, such multiple observations should be of little value in reproducing synoptic patterns at longitudes far from where a scan was taken. Section 6 provides some evidence against this claim.

In all, the TOMS data includes over 100 million observations, so that developing a statistical model for the entire dataset that provides a good description of the joint variations across latitudes, longitudes and seasons would be a monumental conceptual task and fitting such a model an enormous computational task. Gelpke and Künsch (2001) model advective displacement from one day to the next of total column ozone from satellite images in the latitude band 26.5°N to 71.5°N. As such, this work represents a rare attempt to model simultaneously latitudinal and longitudinal variations in total column ozone at high spatial resolution. However, their approach is geared towards analyzing detailed movement of ozone from one day to the next, and not to develop a statistical model for

multiple years of data. Huang, Cressie and Grabosek (2002) propose a model for the whole dataset based on a multiresolution tree-structured approach that is amenable to very fast analysis at the necessary cost of realism in the covariance structure for the total column ozone process. Fang (1996) used a simple advection-diffusion equation driven by white noise to model total column ozone levels at a single latitude, but did not explicitly attempt to model the seasonality nor use formal methods for parameter estimation. The increase in computing power since Fang (1996) opens the possibility of using at least crude approximations of the likelihood function to fit models to all of the data at a single latitude. This paper considers the single latitude 40.5°N, leaving us with $360/1.25 = 288$ possible observations each day. We further restrict to January 1, 1980 to May 6, 1993, because of a higher fraction of missing data in the first year or so of operations. Even with these restrictions, there are still $J = 1{,}404{,}000$ potential ozone observations (288 longtitudes$\times$4875 days), of which slightly under 1% are missing.

Near the equator, there is very little overlap in observations from one orbit to the next, but as one heads poleward, many of the processed observations are in fact based on averages of raw observations from two or more orbits. Thus, it is not a simple matter to identify each processed observation with a specific time. Nevertheless, Fang and Stein (1998) provide spectral analyses showing little influence of the orbital frequency on results at temperate latitudes, so that a reasonable approximation to the times (where time is measured with respect to a fixed location) and locations of the observations is that the jth observation is at longitude $\{\pi - 2\pi(j - 1/2)/288\} \bmod 2\pi$ (in radians) and time $j/288$ (in days) if, in accordance with the gridding scheme used for Level-3 TOMS data, we define the first observation to be at longitude $-179.375°$ and time $1/288$. Near the equator and especially after the Pinatubo eruption, Fang and Stein (1998) do find evidence of artifacts due to the orbital frequency of the satellite, but we will ignore this issue hereafter in this work.

3 Preliminary analyses

All analyses will be done on the natural logarithms of total column ozone (in Dobson units), since this transformation helps somewhat to equalize variation across seasons and makes the process more nearly Gaussian. From here on, we will just say ozone when referring to the natural logarithm of total column ozone.

Stratospheric ozone shows strong seasonal cycles that must be accounted for in any sensible model. To remove seasonal effects from mean ozone levels, a constant term and four harmonics of the annual

cycle was fit via ordinary least squares separately at each of the 288 available longitudes and then subtracted from the observations. Fitting a separate seasonal cycle at each longitude should help to remove real and anomolous differences in ozone values over land and sea (Cuevas, et al. 2001).

Before proceeding with further analyses, it will be helpful to fill in the missing observations. Ideally, one would want to use an approach that accounts for the uncertainty in the unobserved values, but considering that fewer than 1% of the observations in the dataset under study are missing, a simple imputation scheme is unlikely to have much of an effect on our analyses. Thus, we used a simple linear interpolation scheme to fill in any missing observations. Specifically, writing Z_j for $j = 1, \ldots, 1{,}404{,}000$ for the deseasonalized ozone series, if Z_j and Z_{j+k} were observed and $Z_{j+1}, \ldots, Z_{j+k-1}$ were missing, then for $\ell = 1, \ldots, k-1$, the missing observation $Z_{j+\ell}$ was replaced by $\left(1 - \frac{\ell}{k}\right) Z(j) + \frac{\ell}{k} Z(j+k)$.

At this point, one has a deseasonalized ozone series of length 1,404,000, and our main goal in the rest of this work is to model the covariance structure of this series. One major difficulty is that although seasonality has been removed from the mean of the observations, the variation in the data still depends strongly on season, as has been previously noted by Allen and Reck (1997) and Fang and Stein (1998). We provide further evidence of seasonality in the covariance structure here. Define $I_{p,y}(\omega_j)$ to be the periodogram for the time series Z at frequency $\omega_j = 2\pi j/4032$ and the pth two-week period in year y and let $L_{p,y}(\omega_j) = \log I_{p,y}(\omega_j)$. Furthermore, for the frequency band $[\omega_k, \omega_\ell]$, let

$$\overline{L}_p([\omega_k, \omega_\ell]) = \frac{1}{(\ell - k + 1) N_p} \sum_{j=k}^{\ell} \sum_{y=1}^{N_p} L_{p,y}(\omega_j),$$

where N_p is the number of available years for period p, either 13 or 14. For each of five frequency bands, Figure 1 plots deviations of $\overline{L}_p([\omega_k, \omega_\ell])$ from its average over the 26 two-week periods. If Y were stationary, we should see roughly random scatter about 0 for each frequency band. In contrast, we see a strong cyclical pattern at all frequencies, with relatively less power in the late summer and more power in the late winter. If this pattern were the same at all frequencies, that would indicate that the correlation structure does not depend on season and that the inferred nonstationarity could perhaps be removed by a seasonal rescaling. However, the pattern is not the same for all frequency bands: Figure 1 shows a stronger cycle for lower frequencies and a phase about one month ahead of the higher frequencies.

For limited parts of the year, Figure 1 suggests it is reasonable to assume the Z_js are an approximately stationary time series. Figure 2 plots $\frac{1}{14} \sum_{y=1}^{14} I_{1,y}(\omega)$, the average of the 14 periodograms for the first 4032 observations of each year. The most striking feature of this empirical spectrum

is the strong periodicity in the lower frequencies that is highlighted in the top plot of the figure, which reflects the periodicity in the observation pattern, with every 288th observation occurring at the same longitude. It is interesting to note that if we used only one observation for each orbit of the satellite, the frequency of the observations would be reduced by a factor of 288/13.825, which would be roughly like restricting Figure 2 to frequencies up to $(13.825/288)\pi = 0.048\pi$. Visual inspection of Figure 2 indicates that restricting to this frequency range would throw out some real information about fluctuations in the spectrum up to around 0.1π.

There are other features of the covariance structure that we would like to capture in our model. Fang and Stein (1998) note shifts in ozone from one day to the next, always eastward at 40°N, but with a clear seasonal cycle in the size of this shift, with the largest daily shifts generally occurring in the winter. Figure 3 plots the sample autocovariance function corresponding to the empirical spectrum in Figure 2, in which this eastward shift can be seen through the fact that the peaks in the autocovariance function are shifted to the left away from lags of an integer number of days. Thus, fully symmetric covariance functions (Gneiting 2002), which we can define in the present context as covariance functions satisfying $\mathrm{cov}\{Z(\ell,t), Z(\ell',t')\} = \mathrm{cov}\{Z(\ell,t'), Z(\ell',t)\}$ for all ℓ, ℓ', t and t', will not be appropriate. Allen and Reck (1997) observe substantial power in low wavenumbers in total column ozone as well as a noticeable seasonal cycle in this power at more northerly latitudes. Finally, one needs to consider the dependence structure over longer time scales, especially its seasonal structure. For example, Bloomfield, Hurd and Lund (1994) found periodic correlations in monthly stratospheric ozone levels as measured at Arosa, Switzerland (46.8°N) over a 50 year period. We will return to the problem of modeling correlations on longer time scales in Section 8.

4 Model

To describe the ozone process at a single latitude, it is natural to represent it as a Fourier series in longitude with amplitudes evolving in time. Specifically, letting $\mathbb{T}$ be the real line mod 2π, we can write the process $Z(\ell,t)$ on $\mathbb{T} \times \mathbb{R}$ in the form

$$Z(\ell,t) = \sum_{m=0}^{\infty} \{A_m(t)\cos(m\ell) + B_m(t)\sin(m\ell)\}$$

with the A_ms and B_ms stochastic processes. This expression is a partially spectral representation of the process Z. As a special case of results on spectral representations for homogeneous processes on groups (Yaglom, 1961), it is possible to show that for Z to be real-valued, mean 0 (to get other

constant means, just include a nonzero mean in A_0) mean square continuous and weakly stationary on $\mathbb{T} \times \mathbb{R}$, it is necessary and sufficient that we can write

$$\begin{aligned} A_m(t) &= \int_{\mathbb{R}} \cos(\omega t)\widehat{X}_m(d\omega) - \int_{\mathbb{R}} \sin(\omega t)\widehat{Y}_m(d\omega) \quad \text{and} \\ B_m(t) &= \int_{\mathbb{R}} \sin(\omega t)\widehat{X}_m(d\omega) + \int_{\mathbb{R}} \cos(\omega t)\widehat{Y}_m(d\omega), \end{aligned}$$

where the $\widehat{X}_m$s and $\widehat{Y}_m$s are uncorrelated real-valued random measures with orthogonal increments, $E\widehat{X}_m(d\omega)^2 = E\widehat{Y}_m(d\omega)^2 = F_m(d\omega)$ and $\sum_{m=0}^{\infty} F_m(\mathbb{R}) < \infty$. It follows that

$$\text{cov}\{A_m(t), A_m(t')\} = \text{cov}\{B_m(t), B_m(t')\} = \int_{\mathbb{R}} \cos\{\omega(t-t')\}F_m(d\omega),$$

$$\text{cov}\{A_m(t), B_m(t')\} = \int_{\mathbb{R}} \sin\{\omega(t-t')\}F_m(d\omega)$$

and $\{A_m, B_m\}$ and $\{A_n, B_n\}$ are uncorrelated if $m \neq n$. We then get

$$\begin{aligned} \text{cov}\{Z(\ell,t), Z(\ell,t')\} &= \sum_{m=0}^{\infty} \Big[\cos\{m(\ell-\ell')\} \int_{\mathbb{R}} \cos\{\omega(t-t')\}F_m(d\omega) \\ &\quad - \sin\{m(\ell-\ell')\} \int_{\mathbb{R}} \sin\{\omega(t-t')\}F_m(d\omega) \Big]. \end{aligned}$$

If we add the condition that the F_ms are symmetric about 0, then A_m and B_m become uncorrelated and Z is fully symmetric on $\mathbb{T} \times \mathbb{R}$. Since the total column ozone process is clearly not fully symmetric, we might want to consider allowing asymmetry in the F_ms. However, we will instead introduce asymmetries by considering models of the form

$$Z(\ell,t) = \sum_{m=0}^{\infty} \left(A_m(t)\cos[m\{\ell - s_m(t)\}] + B_m(t)\sin[m\{\ell - s_m(t)\}] \right) \tag{1}$$

for deterministic functions s_m. Salby (1987) uses a similar approach to modeling advections across longitudes for trace atmospheric constituents such as ozone. If K_m is the Fourier transform of F_m, we get

$$\text{cov}\{Z(\ell,t), Z(\ell',t')\} = \sum_{m=0}^{\infty} \cos[m\{\ell - \ell' - s_m(t) + s_m(t')\}]K_m(t-t'). \tag{2}$$

The function s_m has a direct interpretation as a rotation for the mth wavenumber and its derivative $S_m(t) = s_m'(t)$ a rotational velocity. Stationarity would require that S_m be constant for each m.

Next, consider what this model for $Z(\ell,t)$ says about the process $\{Z_j\}$. From now on, we will assume that Z_j occurs at time j, so time is measured in steps of 1/288 days. Let f_m be the spectral

density on $(-\pi, \pi]$ corresponding to the covariance function K_m on the integers. Then the spectral density for $\{Z_j\}$ under (2) is

$$f(\omega) = \frac{1}{2} \sum_{m=0}^{\infty} \left\{ f_m\left(\omega + m\left(\frac{2\pi}{288} - S_m\right)\right) + f_m\left(\omega - m\left(\frac{2\pi}{288} - S_m\right)\right)\right\}. \tag{3}$$

Since f determines the covariance function for the observations, the data alone provide no basis for distinguishing between the various f_ms, which is an inherent limitation of data taken by a sun-synchronous satellite. Salby (1982) provides further discussion on the aliasing problems resulting from observations taken from a sun-synchronous satellite. However, by assuming that each f_m is of some simple form, we may be able to get reasonably stable estimates of them. For example, with the f_ms monotonically decreasing on $(0, \pi]$, a sum of the form (3) can match the peaks in the spectrum of $\{Z_j\}$ in Figure 2.

5 Fitting the model season by season

Let us consider fitting (3) to the empirical spectrum in Figure 2. We have several choices to make: what to do about the infinite sum in (3), the form for f_m (or, equivalently, K_m), and the functional dependence of S_m on m. To deal with the infinite sum, we will truncate the sum with an upper limit of M. Modest changes in the value of M do not appear to affect the fit much and we will use $M = 15$ (so 16 terms in the sum), which provides a reasonable compromise between complexity and flexibility. Although it would be easier at this point to specify the spectral densities f_m rather than the covariance functions K_m, when we consider nonstationary models, a time domain specification will be preferable. Following Gneiting and Schlather (2004), we might use $K_m(t) = \alpha_m(1 + |t/\gamma_m|^{2\delta})^{-\nu}$, which is a valid stationary covariance function for all positive α_m, γ_m and ν and all $0 < \delta \leq 1$. As pointed out by Gneiting and Schlather (2004), this covariance function allows one to model the local smoothness of the process and the long-range dependence separately, at least for nondifferentiable processes. We will in fact set $\delta = 1$, not because we believe that A_m is an analytic process on $\mathbb{R}$ as this choice implies, but because it is difficult to estimate both δ and ν and setting $\delta = 1$ yields good fits to the observations. With regards to the S_ms, we will make the simplest possible assumption and take them all equal to S independent of m over each season, although Mote, Holton and Wallace (1991) found evidence for faster propagation at around wavenumbers 6 and 7. Finally, our model for $\{Z_j\}$ includes an independent and identically distributed $N(0, \sigma^2)$ term for each Z_j.

We thus have 35 parameters (16 α_ms and γ_ms, ν, S and σ^2) in the model, which we estimated by maximizing a spectral approximation to the likelihood using the optimization routine nlm in the

statistical package R. Specifically, letting $\omega_j = 2\pi j/4032$ be the jth Fourier frequency, we will assume that $\overline{I}(\omega_j)$, $j = 1, \ldots, 2016$, are independent, $\overline{I}(\omega_j)$ follows a gamma distribution with parameters 14 and $EI(\omega_j)/14$ for $j = 1, \ldots, 2015$ and a gamma distribution with parameters 7 and $EI(\omega_j)/7$ for $j = 2016$. We will not use $j = 0$ in fitting the parameters. This approximation follows from the standard asymptotics for the periodogram (e.g., Theorem 5.2.6, Brillinger 1981) together with the assumption that seasons from different years are independent of each other. To obtain $EI(\omega_j)$ one often uses the approximation $f(\omega_j)$, but here we do not have a convenient expression for this spectral density. Furthermore, when $\nu \leq 1$, f_m (3) is not differentiable at the origin, so that the resulting f is not differentiable in S, which is problematic for a routine like nlm that uses numerical derivatives. Thus, we use the exact formula for the expected value of the periodogram of a stationary process (Priestley 1981, (6.1.37)), which can be computed quickly and exactly using the fast Fourier transform. Figure 2 shows good visual agreement between the fitted model and the empirical spectrum. One can do moderately better by increasing M, but the parameter estimates become increasingly unstable. For example, increasing M to 17 in period 1 adds 4 parameters to the model and increases the loglikelihood by 2.5, which is a quite modest increase considering that it is based on 56,448 observations. Figure 3 shows the empirical and fitted covariance functions for period 1. The agreement between the two is fairly good for the first few days, but the fitted covariances are noticeably off for the peaks near days 3 and 4, being too soon, too large and too spread out, perhaps suggesting that more sophisticated models for space-time asymmetry and long-range correlations are needed.

Figures 4 and 5 give the $\widehat{\alpha}_m$s and $\widehat{\gamma}_m$s for periods 2, 8, 14 and 21 (period 1 is not selected because for this period, as well as for periods 6, 7 and 26, there is considerable instability in the estimate of $\widehat{\alpha}_1$). The general patterns are similar for each of these four seasons. It appears that the patterns in $\widehat{\alpha}_m$ and $\widehat{\gamma}_m$ for $m = 14$ and 15 reflect the model's attempt to fit the high frequency variations in the data rather than any dramatic differences in the temporal evolution of wavenumbers 14 and 15 compared to lower wavenumbers. There is no guarantee that we have found the global maximizers of the likelihood for each season and, indeed, there are instabilities in some parameter estimates beyond those already noted in $\widehat{\alpha}_1$.

Rather than give results for all 35 parameters individually, Figure 6 gives the plots for certain groupings of the parameters for all 26 seasons. The top plot in Figure 6 shows seasonal patterns of the $\widehat{\alpha}_m$s: $\widehat{\alpha}_3$, $\widehat{\alpha}_4$ and $\widehat{\alpha}_5$ as representative of the lower wavenumbers and $\widehat{\alpha}_{14}$ and $\widehat{\alpha}_{15}$ to show the behavior at the highest wavenumbers in the model. Both ranges of m show a clear seasonal pattern,

but the lower wavenumbers show stronger seasonality. Again, this is almost certainly connected to the fitted model using wavenumbers 14 and 15 to track the high frequency behavior in the data rather than any change in seasonality at these wavenumbers. The bottom two plots give the seasonal pattern of $\widehat{\gamma}_3$, $\widehat{\gamma}_4$ $\widehat{\gamma}_5$, $\widehat{\nu}$, $\widehat{S}$ and $\widehat{\sigma}^2$. The fluctuations in the $\widehat{\gamma}_m$s are rather similar to those for $\widehat{\nu}$, which may be natural since both the γ_ms and ν relate to how quickly correlations die out over time. Thus, this similarity is likely at least in part due to statistical correlations rather than due to similarity in the underlying seasonality for the γ_ms and ν. Finally, $\widehat{S}$ shows a clear seasonal pattern, whereas $\widehat{\sigma}^2$ does not. Values for $\widehat{S}$ indicate eastward rotational velocities of around 9° per day in midwinter and under 3° a day in mid to late summer, which qualitatively agree with the findings in Fang and Stein (1998), although there the data were divided into only four seasons, thus flattening out the seasonal pattern considerably.

6 Synoptic prediction

As noted in Section 2, synoptic maps of total column ozone are used for setting the initial conditions of numerical models. Lait (2000) summarizes various methods for producing such maps, including those that are purely data-driven and those that combine data and numerical models, such as 4-D data assimilation (Fisher and Lary, 1995). Because we are here modeling the covariance structure of total column ozone at all longitudes and times, we can, under our estimated model, compute optimal (ignoring the fact that the mean and covariance structures were estimated) linear predictions of total column ozone at any longitude and time, as well as produce mean squared errors for these predictions. In particular, by examining these mean squared errors as a function of observation density, we can examine the claim of Salby (1987) that having more than one observation per scan should not be of much value in producing synoptic maps, especially at locations for which the local time is not near an actual observation time, which is always near local noon for the dataset under consideration here. Lait (2000) examined this question using simulated N_2O fields and various sampling protocols and interpolation methods (none of which involved modeling the covariance structure), finding varying degrees of improvement from using more than one observation per orbit depending on the conditions.

Although we cannot directly examine the effect of using only some fraction of the observations from each orbital pass based on the Level-3 data used here, we can approximate this situation. In particular, $288/21 = 13.714$ is very near the daily orbital frequency of the satellite of 13.825, so using evey 21st observation should be comparable to using one observation per orbit. Consider predicting total column

ozone along latitude 40.5° at local noon on the international date line when it is January 8 west if the line and January 7 east of the line. Using observations from the season January 1–January 14 and the stationary model fit to this season in the preceding section, we can compute mean squared prediction errors of best linear predictors assuming the mean of the process is 0. Figure 7 gives these root mean squared prediction errors under this model using different sampling frequencies. Because these mean squared errors ignore the uncertainty in both the mean and covariance structures, they are likely to be somewhat overly optimistic. However, in light of the large sample sizes, we might hope that the effects of ignoring these uncertainties is modest, and that, in any case, the effect would be similar for different sampling densities, but we have no direct evidence for this claim. Using only every 21st observation leads to strongly oscillating mean squared errors, with local minima not necessarily corresponding to locations at which observations were available on that day. Using every seventh observation greatly dampens these oscillations and does much better overall; using every third observation does nearly as well as using every observation. The quite small root mean squared errors near $-180°$ longitude for the two highest observation frequencies should perhaps not be taken too seriously as they may depend critically on the approximation that the observations are all taken exactly at local noon. One would need to go back to the Level-2 version of the data (Krueger, et al. 1998), which gives ozone measurements at their actual times and locations (so includes observations at essentially the same location taken on successive orbits), to address this question.

Given that it is possible to compute the optimal linear predictors based on two weeks of data using the full density of observations, there seems to be no point in throwing away some of the observations, even if the gains in prediction accuracy are modest. If, on the other hand, when designing an instrument, one has to make a tradeoff between spatial resolution and cost and/or measurement error, it does appear that not much would be lost for purposes of synoptic prediction by having somewhat less dense observations than provided by this TOMS instrument.

Of course, the results in Figure 7 assume that the fitted covariance structure is correct. In fact, in addition to the possible model misfit indicated in Figure 3, the confounding of location and time on a local scale makes it very difficult to distinguish small scale temporal and spatial fluctuations, so our fitted covariance function could be seriously in error. However, it is worthwhile to distinguish between problems with inferring the covariance function and ability to predict synoptic maps once the covariance function is known. What Figure 7 shows is that a model that fits the empirical spectrum of the TOMS data well implies that predictions even around midnight local time can be greatly improved

by using more than one observation per orbit. Before giving this result too much credence, one would want to find further evidence that the fitted covariance structure is appropriate for $Z(\ell, t)$ for all ℓ and t, and not just for the locations in space-time for which TOMS takes observations. By using some combination of other sources of ozone data and theoretical knowledge about the ozone process, it might be possible to address this issue. Another problem that is impossible to address using the TOMS data is the possibility of diurnal cycles in total column ozone. If there were a substantial diurnal cycle, it would have to be estimated using some other source of information in order to account for its impact on a synoptic map.

7 Models for seasonality

From Figure 6, it is apparent that there is a great deal of structure to the seasonality in the covariances, so a model that fits separate parameters for each of the 26 two-week seasons is unnecessarily complex. This section compares various models for the seasonality through their maximized loglikelihoods under the approximation that seasons are statistically independent of each other (in addition to the spectral approximation to the likelihood we are using within seasons). Since there clearly is dependence between at least nearby seasons, we should not take the differences in loglikelihood too seriously, but their orders of magnitude should be indicative of the quality of fit of the various models. The two extreme models we will fit are the model with no seasonality of the form (2) (with 35 parameters) and the "fully seasonal" model that has separate parameters for each season ($35 \times 26 = 910$ parameters). The difference in maximized approximate loglikelihoods between these models is 24,366.

Figure 6 suggests fitting two separate seasonal models to the α_ms, one for $m \leq 13$ and one for $m \geq 14$, and a seasonal model for S. It is not clear from Figure 6 whether seasonal models for the γ_ms, ν and σ^2 are needed. Consider the following parsimonious model for seasonality. Using the subscript j to indicate two-week period j with $j = 1, \ldots, 26$, setting $\zeta = 2\pi \times 14/365.25$, assume

$$\alpha_{mj} = \begin{cases} \alpha_m \exp[T_{\alpha 1} \cos\{\zeta(j-1) + \phi_{\alpha 1}\}], & \text{if } m \leq 13; \\ \alpha_m \exp[T_{\alpha 2} \cos\{\zeta(j-1) + \phi_{\alpha 2}\}], & \text{if } m \geq 14, \end{cases} \tag{4}$$

where $T_{\alpha 1}$ and $T_{\alpha 2}$ are nonnegative and $\phi_{\alpha 1}$ and $\phi_{\alpha 2}$ are in $(-\pi, \pi]$. Forcing the $T_{\alpha j}$s to be nonnegative avoids an identifiability problem in that changing the sign of $T_{\alpha j}$ and adding $\pi \pmod{2\pi}$ to $\phi_{\alpha j}$ yields the same model. Since S need not be positive, assume $S_j = S[1 + T_S \cos\{\zeta(j-1) + \phi_S\}]$. This model has 41 independent parameters and was fit by maximizing the sum of the approximate likelihoods over the 26 two-week periods, yielding a loglikelihood 21,096 greater than the 35-parameter nonseasonal

model and only 3270 smaller than for the 910-parameter fully seasonal model. Adding a common single harmonic seasonal pattern for all β_{mj}s and a single harmonic pattern for the ν_js and σ_j^2s adds six parameters and increases the loglikelihood by only 283 units, whereas allowing only a single cycle for all m for the α_{mj}s removes two parameters but decreases the loglikelihood by 3583. Table 1 gives the estimated seasonality parameters for this model, which shows, as one would expect from Figure 6, much stronger seasonal cycles for the α_{mj}s for $m \leq 13$ than for $m > 13$ and a substantial seasonal cycle for S.

To capture the seasonal influence on the covariance structure, we seek a version of (2) that allows the covariances to change smoothly with season. It is natural to assume that the A_ms and B_ms are periodically correlated (Hurd 1989) with period one year, so that for $D = 288 \times 365.25 = 105{,}192$, $K_m(t + Dj, t' + Dj) = K_m(t, t')$ for any integer j. One way to do this and obtain covariance functions that are guaranteed to be valid is to use the approach of Paciorek and Schervish (2006) to derive nonstationary versions of $K_m(t) = \alpha_m(1+t^2/\gamma_m^2)^{-\nu}$. Specifically, using a simple extension of Paciorek and Schervish (2006) (see Stein 2005), one has that

$$K_m(t,t') = \frac{2^{1/2}\alpha_m(t)\alpha_m(t')\gamma_m(t)^{1/2}\gamma_m(t')^{1/2}}{\{\gamma_m(t)^2 + \gamma_m(t')^2\}^{1/2}} \left[1 + \frac{2^{1/2}(t-t')^2}{\{\gamma_m(t)^2 + \gamma_m(t')^2\}^{1/2}}\right]^{-\{\nu(t)+\nu(t')\}/2} \tag{5}$$

is a valid covariance function for a process on $\mathbb{R}$ for all positive functions α_m, γ_m and ν. In this parameterization, $\mathrm{var}\{A_m(t)\} = \mathrm{var}\{B_m(t)\} = \alpha_m(t)^2$. If we further assume that the functions α_m, γ_m, S and ν are continuous and have period D and all but S are positive, then

$$\mathrm{cov}\{Z(\ell,t), Z(\ell,' t')\} = \sum_{m=0}^{\infty} \cos[m\{\ell - \ell' - s_m(t) + s_m(t')\}] K_m(t,t') \tag{6}$$

with $K_m(t,t')$ as in (5) is a valid periodically correlated version of (2) as long as $\sum_{m=0}^{\infty} \alpha_m(t)^2 < \infty$ for all t. As with the stationary version of this model, we would want to restrict this model considerably before attempting to fit it to data, but it perhaps could serve as a basis for the comprehensive model we seek for all of the data at this latitude.

8 Discussion

Presently, it is not even remotely feasible to calculate the likelihood function exactly for the full dataset under (6). Indeed, even computing all of the covariances is a daunting task, since after taking account of the yearly periodicity in the covariance structure, there are still over 10^{11} different covariances. Thus, we either need to use methods that do not require calculation of all of these covariances or

calculates most of them only approximately, or both. The model reported on in Table 1 exploits both shortcuts: covariances across different two-week periods are ignored, and the process is assumed stationary within any two-week period, so that the covariance matrix is fully specified by its first row. A way to account for covariances across seasons, guarantee a valid covariance structure for the entire dataset, but keep the computations somewhat manageable would be to take $\{\alpha_m\}$, $\{\gamma_m\}$, ν and $\{s_m\}$ in (5) and (6) to be constant within each season. In particular, the covariance matrix of the observations then has Toeplitz blocks of size 4032×4032, which reduces the number of covariances that have to be calculated by a factor of about 2000 and makes possible the use of methods for block Toeplitz matrices to carry out the likelihood calculations. However, full likelihood calculations would still be infeasible at present. A more realistic approach would be to write the 347 two-week seasons as a vector-valued stationary process of length 4032 and assume the vector-valued discrete Fourier transform is complex Gaussian and independent at distinct Fourier frequencies (see Theorem 4.4.1, Brillinger 1981). One would then have to deal with complex covariance matrices of size 347×347, but these matrices would have further exploitable structure due to the annual cycle in the covariance structure. As in Section 5 for a univariate time series, we would not have explicit expressions for the matrix-valued spectral density and could instead calculate these covariances exactly using the fast Fourier transform. Longer seasons would reduce computations but lead to poorer representations of seasonality. Even so, yet further approximations to the likelihood might be necessary.

Before embarking on fitting a single model to the entire time series, one should take account of the fact that variations over longer time scales in total column ozone are influenced by a number of known exogenous factors such as volcanic eruptions, the solar cycle, chlorine loading and QBO (quasibiennial oscillation) (Guillas, et al. 2005), as well as perhaps geopotential height (Peters and Enizian, 1999). Figure 8 plots the two-week averages of the Z_js, which show a sharp jump in 1992, variations in seasonal patterns and a hint of a downward trend. Any modeling of these data on time scales longer than a few months should try to account for these patterns using known exogenous influences before resorting to purely stochastic modeling.

References

Allen, DR, Reck, RA 1997. Daily variations in TOMS total ozone data. *J. Geophys. Res.* **102**: 13,603–13,608.

Bloomfield, P, Hurd, H, Lund, RB 1994. Periodic correlation in stratospheric ozone data. *J. Time Series Anal.* **15**: 127–150.

Brillinger, DR 1981. *Time Series: Data Analysis and Theory*, expanded ed. McGraw-Hill: New York.

Chapman, WA, Cross, MJ, Flower, DA, Peckham, GE, Smith, SD 1974. A spectral analysis of global atmospheric temperature fields observed by the selective chopper radiometer on the Nimbus 4 satellite during the year 1970–1. *Proc. R. Soc. London, Ser. A* **338**: 57–76.

Cuevas, E, Gil, M, Rodriguez, J, Navarro, M, Hoinka, KP 2001. Sea-land total ozone differences from TOMS: GHOST effect. *J. Geophys. Res.* **106**: 27,745–27,755.

Fang, D. 1996. Modeling the Correlation Structure of the TOMS Ozone Data and Lattice Sampling Design for Isotropic Random Fields. Ph. D. thesis, Department of Statistics, University of Chicago.

Fang, D, Stein, ML 1998. Some statistical methods for analyzing the TOMS data. *J. Geophys. Res.* **103**: 26,165–26,182.

Fisher, M, Lary, DJ 1995. Lagrangian 4-dimensional variational data assimilation of chemical-species, *Q. J. R. Meteorol. Soc.* **121**: 1681–1704.

Gelpke, V, Künsch, HR 2001. Estimation of motion from sequences of images: Daily variability of Total Ozone Mapping Spectrometer ozone data. *J. Geophys. Res.* **106**: 11,825–11,834.

Gneiting, T 2002. Nonseparable, stationary covariance functions for space-time data. *J. Am. Statist. Ass.* **97**: 590–600.

Gneiting, T, Schlather, M 2004. Stochastic models that separate fractal dimension and the Hurst effect. *SIAM Review* **46**: 269–282.

Guillas, S, Tiao, GC, Wuebbles, DJ, Zubrow, A 2005. Statistical diagnostic and correction of a chemistry-transport model for the prediction of total column ozone. *Atmos. Chem. Phys. Discuss.* **5**: 10421-10453.

Hayashi, Y 1983. Modified methods of estimating space-time spectra from polar-orbiting satellite data, Part I and Part II, *J. Meteorol. Soc. Jpn.* **61**: 254–262.

Huang, HC, Cressie, N, Gabrosek, J 2002. Fast, resolution-consistent spatial prediction of global processes from satellite data. *J. Comput. Graph. Statist.* **11**: 63–88.

Hurd, HL 1989. Nonparametric time series analysis for periodically correlated processes. *IEEE Trans. Info. Theo.* **35**: 350–359.

Krueger, AJ, Bhartia, PK, McPeters, RD, Herman, JR, Wellemeyer, CG, Jaross, G, Seftor, CJ, Torres, O, Labow, G, Byerly, W, Taylor, SL, Swissler, T, Cebual RP 1998. *ADEOS Total Ozone Mapping Spectrometer (TOMS) Data Products User's Guide*. National Aeronautics and Space Administration: Greenbelt, MD. Available at <toms.gsfc.nasa.gov/datainfo/adeos_userguide.pdf>.

Lait, LR 2000. Effects of satellite scanning configurations on derived gridded fields. *J. Geophys. Res.* **105**: 9063–9074.

Lait, LR, and Stanford, JL 1988. Applications of asynoptic space-time Fourier transform methods to scanning satellite measurements. *J. Atmos. Sci.* **45**: 3784–3799.

Liu, X, Newchurch, MJ, Kim, JH 2003. Occurrence of ozone anomalies over cloudy areas in TOMS version-7 level-2 data. *Atmos. Chem. Phys.* **3**: 1113–1129.

Mote, PW, Holton, JR, Wallace, JM 1991. Variability in total ozone associated with baroclinic wave. *J. Atmos. Sci.* **48**: 1900–1903.

Paciorek, CJ, Schervish, M 2006. Spatial modelling using a new class of nonstationary covariance function. In press, *Environmetrics*. DOI: 10.1002/env.784.

Peters. D, Entzian, G 1999. Longitude-dependent decadal changes of total ozone in boreal winter months during 1979–1992. *J. Climate* **12**: 1038–1048.

Priestley, MB 1981. *Spectral Analysis and Time Series*, vol. 1. Academic Press: London.

Salby, ML 1982. Sampling theory for asynoptic satellite observations. *J. Atmos. Sci.* **39**: 2577–2600.

Salby, ML 1987. Irregular and diurnal variability in asynoptic measurements of stratospheric trace series. *J. Geophys. Res.* **92**: 14,781–14,805.

Stein, ML 2005. Nonstationary spatial covariance functions. Unpublished technical report. Available at <galton.uchicago.edu/~cises/research/cises-tr21.pdf>.

Yaglom, AM 1961. Second-order homogeneous random fields. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 2, Neyman J (ed); University of California Press: Berkeley, CA; 593–622.

Table 1. Seasonality parameters. Reported phases are estimates (in radians) divided by π and, by definition, fall in $(-1, 1]$.

	$\alpha\,(m \leq 13)$	$\alpha\,(m \geq 14)$	S
Scale	0.9495	0.7268	0.4608
Phase	−0.1779	−0.4481	0.8817

Figure captions

Figure 1. Seasonal deviations in logperiodogram from annual average for selected frequency bands. Band 1–50 (i.e., $[\omega_1, \omega_{50}]$) given by black solid line, band 51–150 by black dashed line, band 151–300 by black dotted line, band 301–600 by gray solid line, band 601-2016 by gray dashed line.

Figure 2. Empirical (gray symbols) and fitted (black curve) spectra for period January 1–14, split into two frequency bands to highlight the strong periodicity at lower frequencies. The empirical spectrum is the average of the 14 periodograms for January 1–14 for 1980–1993.

Figure 3. Emprical (black dashed) and fitted (gray curve) correlation functions for period January 1–14.

Figure 4. Estimates of α_m for $m = 0, \ldots, 15$ for four two-week periods. Solid curve is period 2, dashed is period 8, dotted is period 14 and dotdash is period 21.

Figure 5. Estimates of γ_m for $m = 0, \ldots, 15$ for four two-week periods. Line types as in previous figure.

Figure 6. Estimates for the 26 two-week periods. Upper plot gives $\widehat{\alpha}_m$ for $m = 3, 4, 5$ (gray solid, dashed and dotted, respectively) and $m = 14, 15$ (black dashed and dotted, respectively). Middle plot gives $\widehat{\gamma}_m$ for $m = 3, 4, 5$ (gray solid, dashed and dotted, respectively) and $\widehat{\nu}$ (black dashed). Bottom plot gives eastward movement $-\widehat{S}$ (solid), $\widehat{\sigma}^2$ (dashed).

Figure 7. Root mean square prediction errors for a day in the middle of the Jan. 1–Jan. 14 season as a function of observation density. Solid gray curve uses every observation, black dashed every third observation, gray dashed-dotted every seventh, black dotted every 21st, $\times$s indicate longitudes of available observations on day of prediction when use every 21st observation. Results for positive longitudes are omitted because of symmetry. For comparison, the estimated standard deviation for any one observation based on the fitted model is 0.096.

Figure 8. Two-week averages of adjusted log ozone values at latitude 40.5° N. Dashed vertical lines drawn to highlight seasonal variations.

Figure 1

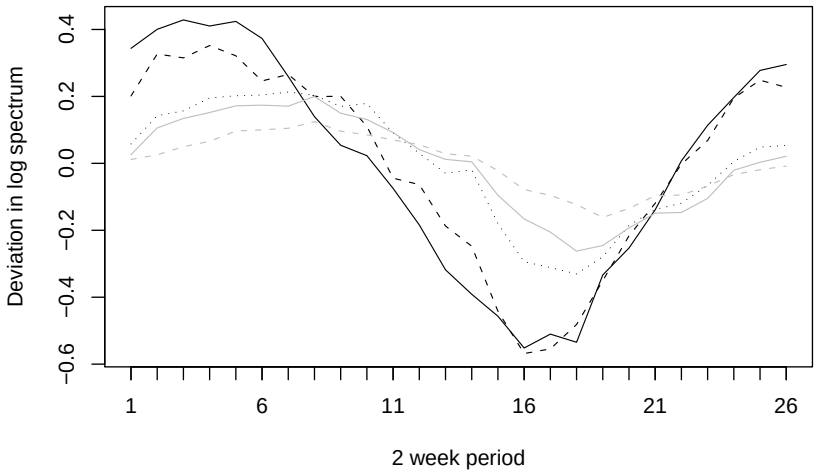
Deviation in log spectrum
0.4
0.2
0.0
−0.2
−0.4
−0.6
1
6
11
16
21
26
2 week period

Figure 2

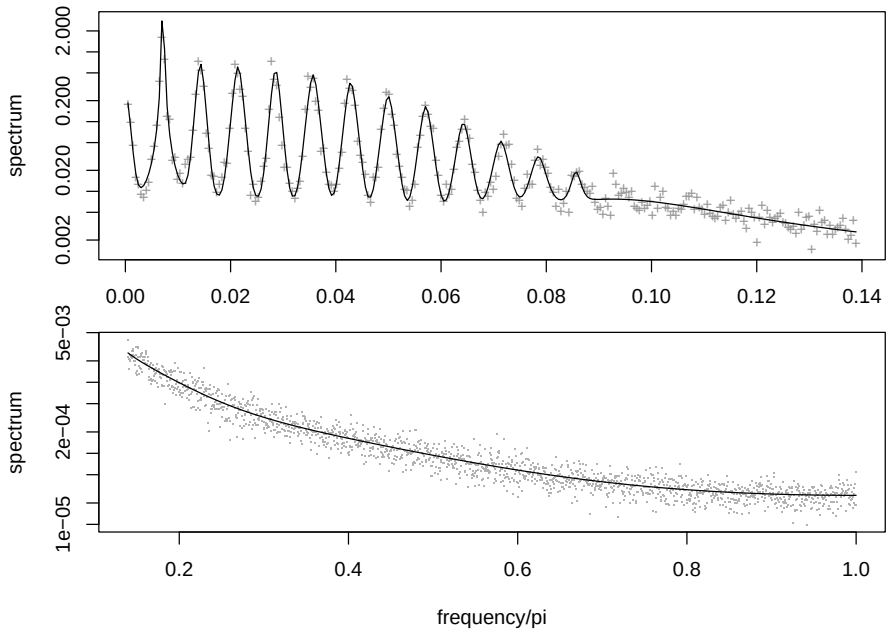
spectrum
2.000
0.200
0.020
0.002
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
spectrum
5e−03
2e−04
1e−05
0.2
0.4
0.6
0.8
1.0
frequency/pi

Figure 3

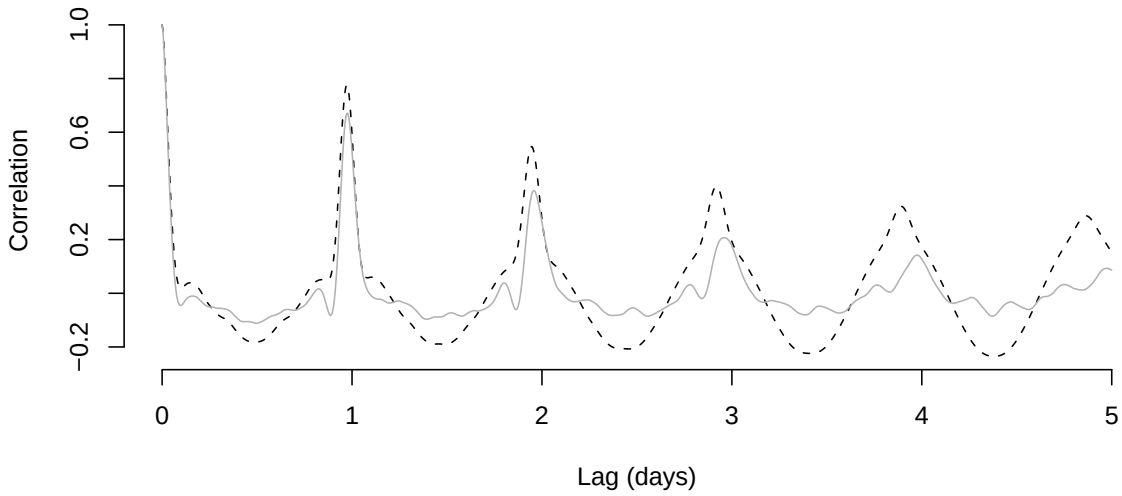

Figure 4

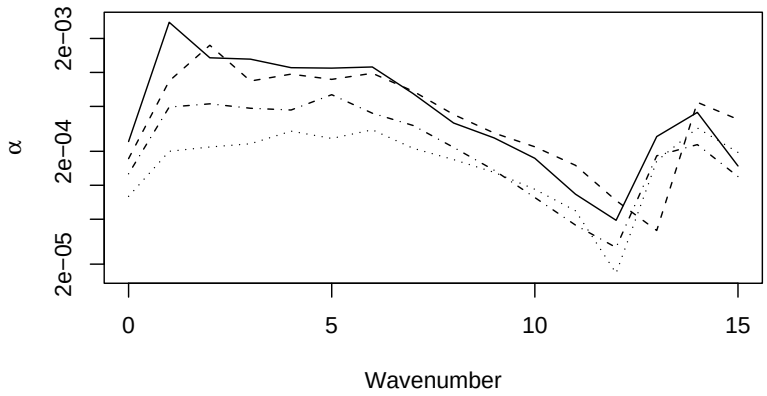

Figure 5

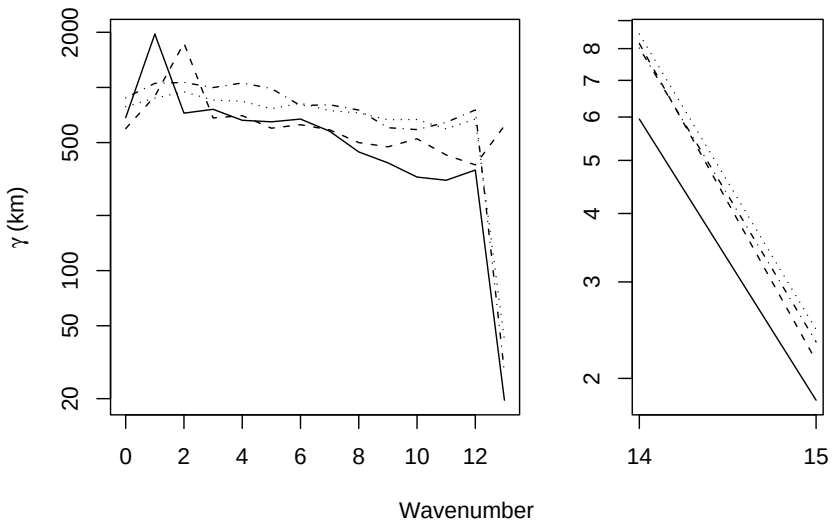

Figure 6

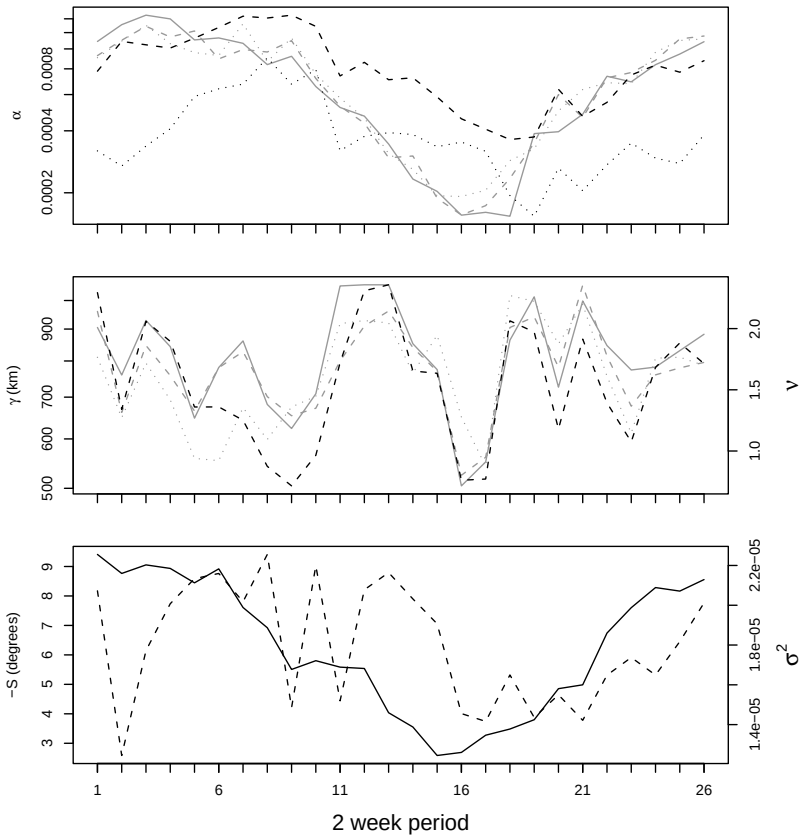
α
0.0002
0.0004
0.0008
γ (km)
500
600
700
900
ν
1.0
1.5
2.0
−S (degrees)
3
4
5
6
7
8
9
σ²
1.4e−05
1.8e−05
2.2e−05
1
6
11
16
21
26
2 week period

Figure 7

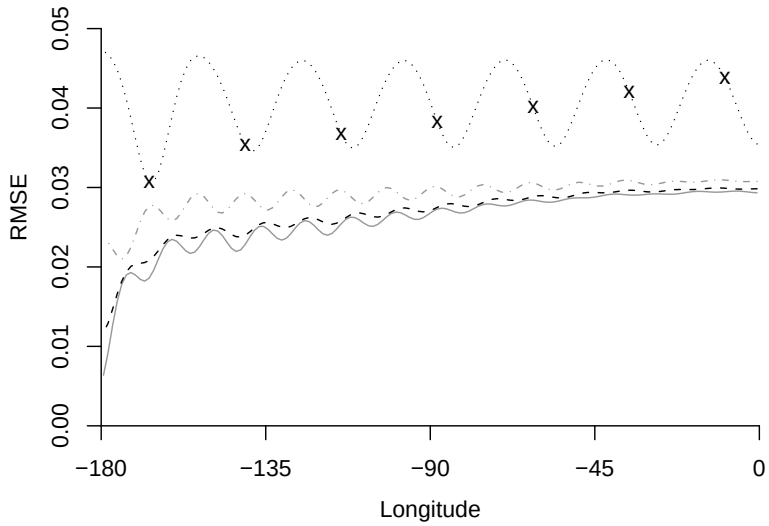

Figure 8