Dynamic Harmonic Regression and irregular sampling; avoiding pre-processing and minimising modelling assumptions

David A. Mindham, Wlodzimierz Tych
Lancaster Environment Centre,
Lancaster University, Lancaster LA1 4YQ, U.K.

Abstract

Many environmental time-series measurements are characterised by irregular sampling. A significant improvement of the Dynamic Harmonic Regression (DHR) modelling technique to accommodate irregular sampled time-series, without the need for data pre-processing, has been developed. Taylor's series is used to obtain the time-step state increments, modifying the transition equation matrices. This allows the user to avoid artefacts arising and insertion of assumptions from interpolation and regularisation of the data to a regular time-base and makes DHR more consistent with the Data-Based Mechanistic approach to modelling environmental systems. The new technique implemented as a Matlab package has been tested on demanding simulated data-sets and demonstrated on various environmental time-series data with significantly varying sampling times. The results have been compared with standard DHR, where possible, and the method reduces analysis time and produces unambiguous results (by removing the need for pre-processing – always based on assumptions) based only on the observed environmental data.

1.0 Introduction

In analysing environmental data and modelling environmental processes there is a significant need to identify and estimate trends, cycles, and seasonal components. Dynamic Harmonic Regression (DHR) provides a cogent analytical tool to generate such results. It is, however, firmly based in the classical time series analysis domain and relies on the data being sampled at specific intervals. However, in many disciplines dealing with the natural environment, data-sets are not sampled at regular intervals. Presented here is a significant update to DHR allowing direct use of irregularly sampled time series data in estimation of trends, cycles and seasonal components. In addition, there is a distinct need for such a method to be accompanied by software that is easy to use and with results directly interpretable in the terms of the specific discipline, be it climatology, hydrology or environmental chemistry, to cite a few of the disciplines where these methods have been successfully applied by the authors.

Dynamic Harmonic Regression is a nonstationary time-series analysis approach used to identify trends, seasonal, cyclical and irregular components within a state space framework (Young, 1989). The DHR method is implemented in the CAPTAIN Toolbox for Matlab and has been used extensively by many researchers (Young et al., 1999, Taylor et al., 2007). The DHR methodology has a wide range of applications, and is particularly useful in analysing environmental data; such as atmospheric pollutants (Becker et al., 2006, Venier et al., 2012) where, importantly, it is cited as a recommended method in the 2011 UNEP Air Report in the Persistent Organic Pollutants section (UNEP, 2011). Other significant applications include paleoclimatology data based on isotope dating (Smith et al., 2016), impacts on catchment water balance (Chappell and Tych, 2012), groundwater-surface water fluxes (Keery et al., 2007), geomorphology (Carling et al., 2005), water quality cycles (Halliday et al., 2013), or solar irradiation forecasting (Trapero et al., 2015), but also in forecasting of phone-call numbers within the call-centre context (Tych et al., 2006), medicine (Sofianopoulou et al., 2017) and

finance (Bhar, 2010). However, when data are irregularly sampled, the existing DHR and related methods cannot be applied directly. For instance, paleoclimatic data-series from core samples and speleothems are interpolated onto a fixed time-base prior to analysis (e.g. Smith et al., 2016). Historic water quality data, geomorphological data and atmospheric chemistry data (e.g. Becker et al., 2006, Carling et al., 2005) are treated in the same way using prior processing.

The problem is that the original state-space filtering-based DHR cannot handle irregular sampling without applying resampling techniques making the time sampling uniform prior to the analysis. As useful as resampling is, it is still manipulation of the observed data and leads to increased uncertainties in model outputs and to potential artefacts resulting from the interpolation techniques applied, such as aliasing (Chappell et al., 2017) or spectral features of the approximation functional base applied in the interpolation process. Importantly, where the actual samples become sparser it can lead to 'false certainty' - introducing interpolated samples where there are no data available. The uncertainty estimates then become tainted, usually unduly lower. Conversely, where samples are denser it can lead to a removal of information, also leading to increasing uncertainty (fewer samples - less averaging) and potentially losing information in the upper part of the signal spectrum. Common anti-aliasing methods for down-sampling (such as low-pass filtering) will have the latter problem.

With interpolation, resampling is a step away from the Data Based Mechanistic approach (DBM, Young, 1999) to modelling and data analysis which DHR is designed to be consistent; allowing the observed data to tell us about the systems prior to process interpretation. This is because interpolation always has an underlying model or assumptions, which may form an introject affecting the data.

It has to be pointed out that the developed algorithm is not aimed purely at dealing with irregular sampling, but at augmenting the existing DHR model which has proven to be highly effective and widely used in Environmental Science due to the natural interpretation of its object and of the model components, as well as the inherent stochastic information provided by it.

Analysing irregularly sampled time series data has a large body of literature addressing it. Irregularly sampled Auto-Regression is one of them. Broersen et al. (2004) derive a method for handling AR models with missing samples (a very specific and limited form of irregular sampling – with missing samples the sampling is regular). In general, AR and other methods relying on solving stochastic equations (such as Brockwell's (2001) Levy process driven approach) are not directly comparable with the proposed technique because they are usually much more general, and so rely on additional mathematical analysis and assumptions in every specific application. Other approaches to irregularly sampled time series address whole spectrum estimation (such as irregularly spaced approaches to Fourier estimation, such as O'Toole et al. (2007) or wavelet-based approaches of e.g. Mathias et al., 2004), which is exactly what is avoided here in order to reduce the uncertainty of results. The reduction of uncertainty in the presented approach is achieved through minimising the number of spectral components that are estimated, to only the dominant periodicities. Various machine learning approaches tend to require high data volumes and suffer from difficulty with obtaining justified uncertainty estimates. We work with often expensive to obtain environmental data sets of necessarily limited length, and normally address univariate time series, so direct comparison with most published machine learning approaches is not easily achieved.

The term "arbitrary sampling" is introduced in the specific DHR context and used to describe how the irregular sampled time-series is used within the irregular sampled DHR technique. The temporal distance between each sample in the irregularly sampled time-series is stored as a 1x(n-1) vector

complementing the irregularly sampled time-series itself. The term 'temporal distance' is used here deliberately to highlight the possibility of using this technique in analysis of spatial series, as for example in Carling et al. (2005) where DHR was used on regularised spatial data to analyse the pool-riffle sequence in river geomorphology. The arbitrary sampling processing uses both series – the measured values and their sample times. This approach can be used also for sparse regularly sampled series with many missing values, where the time index for the missing values are removed, creating an irregular sampled time-series. The creation of the temporal distance vector is the only pre-processing required for the updated DHR and for the purposes of differentiating between the current DHR methodology and the proposed updated methodology, the latter will be referred to as 'Arbitrary Sampled Dynamic Harmonic Regression' (ASDHR).

The update implements an arbitrary sampling technique in the Kalman Filter (KF) and Fixed Interval Smoother (FIS) algorithms. While the irregular sampling has been previously used with Kalman Filtering (e.g. Li et al., 2008), the FIS algorithm implementation here is a novel element, not used elsewhere (except for Mindham et al., 2018) and necessary for the use of ASDHR. Overall the Arbitrary sampling technique eliminates the need for any pre-analysis or resampling and puts DHR back in line with the DBM approach by not inserting any assumptions or artefacts into the observed data.

The aim of this paper is to introduce the arbitrary sampling technique and to demonstrate the benefits of ASDHR for analysis of environmental data sets, which so often are irregularly sampled or contain numerous missing values. This is achieved by:

- Providing a brief background to DHR and then introduce the arbitrary sampling methodology (Section 2.0)
- Demonstrating the capability, benefits and necessity of ASDHR when using environmental data:
    - Paleo-climatology (Smith et al., 2016) – comparing ASDHR and DHR outputs to demonstrate the arbitrary sampling capability (Section 3.1).
    - Persistent organic pollutants (Becker et al., 2006) – demonstrating the necessity for extending DHR to accommodate irregular sampled time-series (Section 3.2, 3.3) especially for noisy data series.
    - Forecasting Atmospheric CO_2 – introducing and demonstrating 'arbitrary forecasting', the ability to forecast at arbitrary points into the future with different sampling times to the observed data (Section 3.4).
- Evaluation of ASDHR robustness to data sparseness and observational noise (Section 4.0)

2.0 Dynamic Harmonic Regression

The DHR model assumes that the observable variable of a system is composed of four components (1); trend (T), sustained cyclical (C) with period different to the seasonality, seasonal (S) and white noise (e) (Young et al., 1993).

$$y_t = T_t + C_t + S_t + e_t \quad (1)$$

The measured values of *y* are the output (observations) series of a system of stochastic state space equations, which can then be broken down to allow for estimation of the four components.

T_t is the trend component, which can be considered a stochastic time-varying 'intercept' parameter and is interpreted spectrally as a zero frequency term (*i=0*, where ω_0 or f_0 = 0), in practice -

occupying the lowest part of the spectrum, and modelled as Integrated (or Smoothed) Random Walk (see Young et al., 1999) with states termed level and slope of the trend.

The seasonal component S_t is defined as:

$$S_t = \sum_{i=1}^{R_s} \{a_{i,t} \cos(\omega_i t) + b_{i,t} \sin(\omega_i t)\} \quad (2)$$

where $a_{i,t}$ and $b_{i,t}$ are stochastic Time-Varying Parameters (TVP) and ω_i are the fundamental and harmonic frequencies associated with the seasonality in the series ($i=1,2,\ldots,R_s$).

$$C_t = \sum_{i=1}^{R_c} \{\alpha_{i,t} \cos(f_i t) + \beta_{i,t} \sin(f_i t)\} \quad (3)$$

where $\alpha_{i,t}$ and $\beta_{i,t}$ are stochastic TVP and f_i are the frequencies associated with the longer cyclical component ($i=1,2,\ldots,R_c$). The cyclic component C_t has an identical definition to the seasonal and is isolated here to allow for a different physical interpretation.

Whereas the white noise component e_t is the remaining information after the other 3 components have been removed from y (i.e. model residuals). Note that the full Unobserved Components Model (Young et al., 1999) also incorporates the Irregular component, here omitted for simplicity.

Typically, the TVP ($a_{i,t}$, $b_{i,t}$, $\alpha_{i,t}$, $\beta_{i,t}$ and both T_t states) are defined by a two dimensional state vector $x_{i,t} = \left[l_{i,t}, d_{i,t}\right]^T$, where $l_{i,t}$ and $d_{i,t}$ are, respectively, the changing level and slope of the associated TVP. The stochastic evolution of each $x_{i,t}$ is assumed to be described by a generalised random walk process (4).

$$x_{i,t} = F_i x_{i,t-1} + G_i \eta_{i,t-1} \quad i = 1,2,\ldots,R \quad (4)$$

where, $R = 1+R_c+R_s$ and F and G defined in their time-varying form in (7a) and (7b) respectively (see also Young et al., 1999 for fixed form).

2.1 State Space and Observation Equations

The state space model is constructed by aggregation of the subsystem matrices defined in (2) and is defined in Young et al. (1999). However, both the state transition and noise-input matrices are fixed and thus can only work for uniformly sampled data, hence the need to apply regularisation techniques on irregularly sampled data. The method proposed here replaces these fixed matrices with time-step dependent ones, where the values depend on the time between each sample; thus, allowing them to work for irregularly sampled data.

For the rest of the paper, the temporal positioning of samples (t) at regular intervals Δt is replaced with the arbitrary positioning of samples (Δt_k), where k – the sample number - keeps the temporal order.

If $y^{\nu}(\boldsymbol{y}_k)$ is the ν^{th} derivative of $y(\boldsymbol{y}_k)$, and the form of the function $y(.)$ is not specified, a data point distant from $\boldsymbol{y}_k$ provides very little information about $y(\boldsymbol{y}_k)$. Using the local polynomial modelling reasoning (*e.g.* Fan and Gijbels, 1996) only the local data points in the vicinity of $\boldsymbol{y}_k$ are used. Assuming $y(\boldsymbol{y}_k)$ has the $(q+1)^{\text{th}}$ derivative at the point $\boldsymbol{y}_k$, then following Taylor's expansion for $\boldsymbol{y}$ in the local neighbourhood of $\boldsymbol{y}_k$ we have:

$$y(\boldsymbol{y}) = y(\boldsymbol{y}_k) + y'(\boldsymbol{y}_k)(\boldsymbol{y} - \boldsymbol{y}_k) + \frac{y''(\boldsymbol{y}_k)}{2!}(\boldsymbol{y} - \boldsymbol{y}_k)^2 + \cdots + \frac{y^q(\boldsymbol{y}_k)}{q!}(\boldsymbol{y} - \boldsymbol{y}_k)^q \quad (5)$$

If the value of y and its derivatives are known at the t^{th} point as $\boldsymbol{x_k} = [y(\varsigma_k)\ y'(\varsigma_k)\ y''(\varsigma_k)\ \cdots\ y^{(q)}(\varsigma_k)]^{\mathsf{T}}$ and the highest derivative of $y(\boldsymbol{y})$ with respect to $\boldsymbol{y}$ with $\boldsymbol{y}_k = y(\varsigma_k)$ and $y^{(q+1)}(\varsigma_k) = \eta_k$, where

$\eta_k \sim \mathcal{N}(0, \sigma_\eta^2)$ and ς_k is the approximation point (knot) at sample k, then Taylor's expansion (3) can be applied in the local neighbourhood of ς_k for all derivatives of y resulting in the GRW model with state space (6a) and observation (6b) equations with the now time-varying state transition (in 7a) and system noise-input (in 7b) matrices:

$$x_k = \boldsymbol{F}_k x_{k-1} + \boldsymbol{G}_k \eta_{k-1} \quad \text{(6a)}$$

$$\boldsymbol{y}_k = \boldsymbol{H}_k^T x_k + e_k \quad \text{(6b)}$$

where H_k is the observation matrix (dimension of nxR for RW trend and RW harmonics' amplitudes). Observation equation (6b) implements the regressive structure of DHR, with x_k being the estimated amplitudes of harmonics (or trend levels) for each *k* and their corresponding elements of H_k contain the i-th's harmonic values ($\cos(\omega_i t_k)$ or $\sin(\omega_i t_k)$) or ones for the trend level.

$$F_k = \begin{bmatrix} \alpha & \Delta_k & \frac{\Delta_k^2}{2!} & \cdots & \frac{\Delta_k^q}{q!} \\ 0 & 1 & \Delta_k & & \frac{\Delta_k^{q-1}}{(q-1)!} \\ 0 & 0 & 1 & & \frac{\Delta_k^{q-2}}{(q-2)!} \\ & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \text{(7a)}, \quad G_k = \begin{bmatrix} \frac{\Delta_k^{q+1}}{(q+1)!} & \frac{\Delta_k^q}{q!} & \frac{\Delta_k^{q-1}}{(q-1)!} \ldots \Delta_k \end{bmatrix}^T \text{(7b)}$$

where, $\Delta_k = \varsigma_k - \varsigma_{k-1}$ is the temporal distance between the temporal samples number k and k-1 (time difference between the knots ς_k and ς_{k-1}). When $\alpha = 1$ and $q = 0,1,2\ldots$ this equation describes, respectively, random walk, integrated random walk, double integrated random walk, etc. When α ranges between 0 and 1 referring to smoothed random walk with orders q as above.

The choice of Random Walk order depends on the application. Where variation is expected q of 1 or higher is used naturally depending on the shape of the trend, where no assumptions are made and suspicion of stationarity needs to be evaluated, the simple RW model ($q=0$) is usually more appropriate. In practice the analysis starts with q of 0 for the harmonic components and 1 for the trend.

2.2 Algorithmic Considerations

To reiterate, the term 'arbitrary sampling' is used because the new state space equations (section 2.1) do not constrain the time between each sample to be fixed, as long as they are in temporal order; the regularity or amount of irregularity in the sampling process does not impact the modelling process so long as the temporal positions of each sample are known and the finite Taylor expansion (5) is a satisfactory approximant of y.

To keep the algorithm numerically well defined (keeping the spectrum of the transition matrix "sensible" as its inverse is used in the smoothing algorithm), the temporal distances between each sample (Δ_k) may need scaling to keep the 'majority' or average sampling rate close to one. This is equivalent to choosing the time unit suitable for the analysis, e.g. if the sampling process is typically once a week, then Δ_k provided in days needs to be divided by seven. In formal terms, as Δ_k affects the spectrum of the transition matrix F_k and its invertibility (condition), the time must be expressed in the units that will not cause poorly conditioned F_k for any (or at least for very few) steps k. In the state space matrices (section 2.1) a Δ_k of 1 for all k implies regular sampling as a special case.

2.3 Variance Intervention

Amongst the numerous algorithmic advantages of the Stochastic State-Space techniques, such as the ease of forecasting, smoothing and interpolation, including arbitrary times between the existing samples, the variance intervention technique seems particularly well suited to environmental data analysis and modelling. Very often the researcher is looking for confirmation or detection of a discrete change in the system. Variance intervention technique has been introduced by Box and Tiao (1975) and in the context of stochastic state space models it was developed and evaluated by Young and Ng (1989). With regular DHR, intervention points are used to account for abrupt changes in the data, such as a sharp calibration change or a shift in environmental system behaviour (e.g., Chappell and Tych, 2012). It can be used to model and evaluate the potential for specific breaks in the time series, whether background level, slope changes or sudden amplitude changes of the harmonic components in the dynamic harmonic regression context. Without intervention points any abrupt changes are smoothed in the model estimate and give poorer models that are not true to the system, not reflecting the mechanisms governing the observed processes and so not consistent with the DBM approach. In Bayesian terms interventions amount to introducing diffused priors at the intervention *a-priori* step, causing the recursive estimator to "doubt" the current estimate by increasing the covariance matrix significantly.

Introducing intervention points requires either assumptions or knowledge of the time of the change. Alternatively, a search for a significant parametric change may be made using a sliding intervention technique, as applied in Chappell and Tych (2012) to detect discrete changes in streamflow and evaporation records due to forest cover change, or to other such interventions.

This advantage is not lost with the introduction of the arbitrary sampling technique, as one of the examples below demonstrates.

2.4 Period Identification

In many environmental data series there is a need for identification of dominant periodicities, as these are likely to indicate the phenomena active or dominant in the processes generating the time series. Spectral estimators such as FFT and the various families of parametric spectral estimators (from Burg's to wavelets), while commonly used, are (a) sensitive to noise, especially coloured noise, and (b) their uncertainty is very high: as Fisher (1929) has shown, the uncertainty of spectral estimators is of the same order as their magnitude. These issues are aggravated for noisy processes within time varying spectra, for the simple reason of the number of estimated values of the spectral characteristics being of a similar magnitude to that of the number of data points. So, while using the standard methods, we are getting a picture for a range of frequencies, this picture is highly uncertain for spectral estimators. With DHR and ASDHR, one periodicity (a handful of harmonic frequencies) is analysed in each step of the periodicity sweep, as we show below. The powerful handling of uncertainty by the Kalman Filter and Fixed Interval Smoother allows for a significant improvement of detection and estimation process, and in addition, importantly and nearly uniquely, provides an uncertainty estimate of the identified periodicities.

The periodicity identification method used for ASDHR involves scanning through a predetermined selection of periods and selecting the most statistically likely period(s). This is relatively time consuming for wider sweeps, but the search could be optimised using variations of common search algorithms. Computation time is arguably not such a critical issue, as (a) this is an off-line process, and (b) with high speed modern computers it is not significant.

The question of identification criterion for spectral peaks is quite critical. The standard R^2, being the proportion of variance of observed data explained by the model (8),

$$R^2 = 1 - \frac{\|e_k\|^2}{\|y_k\|^2} \quad (8)$$

with e_k being the model residuals, could be used as the statistical measure, but this was found to be less reliable, as many periods were found to have similar R^2 values leading to poor sensitivity, especially in trend-dominated series, and so the statistically best period was hard to distinguish from other significant periods. The same approach will apply to multi-periodic signals thanks to the orthogonality of the harmonic components.

Introduced here is an analogue of the standard R^2, a new measure easily described as the proportion of data explained by the seasonal or cyclic component (9).

$$R_s^2 = \frac{\|S_k\|^2}{\|y_k - T_k\|^2} \quad (9)$$

Where, S_k is the estimated seasonal component (section 2.0). Its quadratic norm (variance) is compared with that of the detrended data term in the denominator of (9). Environmental data often have a significant slow (or trend) component, which dominates the standard R^2, while R_s^2 focuses on the detrended data and seasonal component. Note that in the process of identification the trend is estimated together with the seasonal component for each case, so there is no danger of introducing artefacts due to this procedure, effectively a spectral decomposition. In addition, (9) provides a standardised measure of a relative strength of periodicity, comparable across different time series.

2.5 Noise Variance Ratio (NVR)

The introduction of Taylor's expansion to the GRW models indirectly influences the selection of hyperparameters by introducing the sampling rate directly into the state transition equation. For the original DHR spectral model fit was used, as it was relatively easy to formulaically express the DHR spectrum and compare it to the AR spectrum of the data. This becomes more complex and ambiguous for irregularly sampled data. However, NVRs also carry the interpretation of time scaling of the model (spectral boundary of the low pass filter interpretation of the process, as explained in Young, 1999.) In this case, with the challenging application examples we found that choosing the NVRs needs to reflect the time scale of the modelled process, rather than the best fit in any particular sense. The latter would have been arbitrary and would introduce additional assumptions into the modelling process.

3.0 Demonstrating ASDHR on environmental time-series

The examples compare, where possible, the proposed ASDHR methodology with the original DHR methodology. Direct comparison is often difficult as both methods operate on different data sets, DHR works with regularised and sometimes pre-filtered data whereas ASDHR works with the raw or 'unedited', observed data. In terms of the DBM approach or philosophy it should already be apparent that ASDHR is a more appropriate tool for analysing environmental systems.

The first example (3.1) has more comparable data sets (interpolated and raw data) and is a good demonstration of a working ASDHR methodology that is comparable to DHR. The second example (3.2) demonstrates the need for ASDHR to achieve data analysis that is in line with the DBM philosophy, i.e. the data tells the story and not the assumptions used to manipulate the data for DHR analysis. The third example (3.3) introduces a new type of forecasting, termed 'arbitrary forecasting', that allows predictions to occur at chosen points in the future, not just from the end of the observed time-series, and at different sampling rates and points to that of the observed time-series.

3.1 Paleo-climatology example

Analysing patterns such as trends and cycles in paleoclimatic data is a common theme in this discipline. The example we used was a typical one and previously published in Nature: Scientific Reports indicating the importance of the problem for the community. A carbonate oxygen isotope ($\delta^{18}O$) record derived from speleothems contained within Cueva de Asiul situated in the Matienzo depression (Cantabria), North Spain, was used to reconstruct the precipitation delivery to northern Spain during the last 12100 years using DHR analysis to find the trends and periods (for full details and the analysis see Smith et al., 2016).

Here the DHR analysis, in terms of trend and harmonic amplitudes, is compared with the proposed ASDHR. Similar results are expected as both methods are operating on the same data, but the result specifics should be different due to the differences in the data pre-processing and analysis algorithms.

3.1.2 Data Pre-processing

With current DHR the data needs pre-processing to make uniform the sampling rate and this involves linear interpolation followed by removing the interpolated samples where there are gaps in the observed data to avoid introduction of artefacts. In this case (analysed in Smith et al., 2016 by the corresponding author) a highly irregularly sampled data-series of 1919 values is reduced to a uniformly sampled data-series of 815 (62 of which are missing values, mainly due to a single large gap in the data-series), which is a significant data loss, even if the key characteristics of the data are preserved (Figure 1).

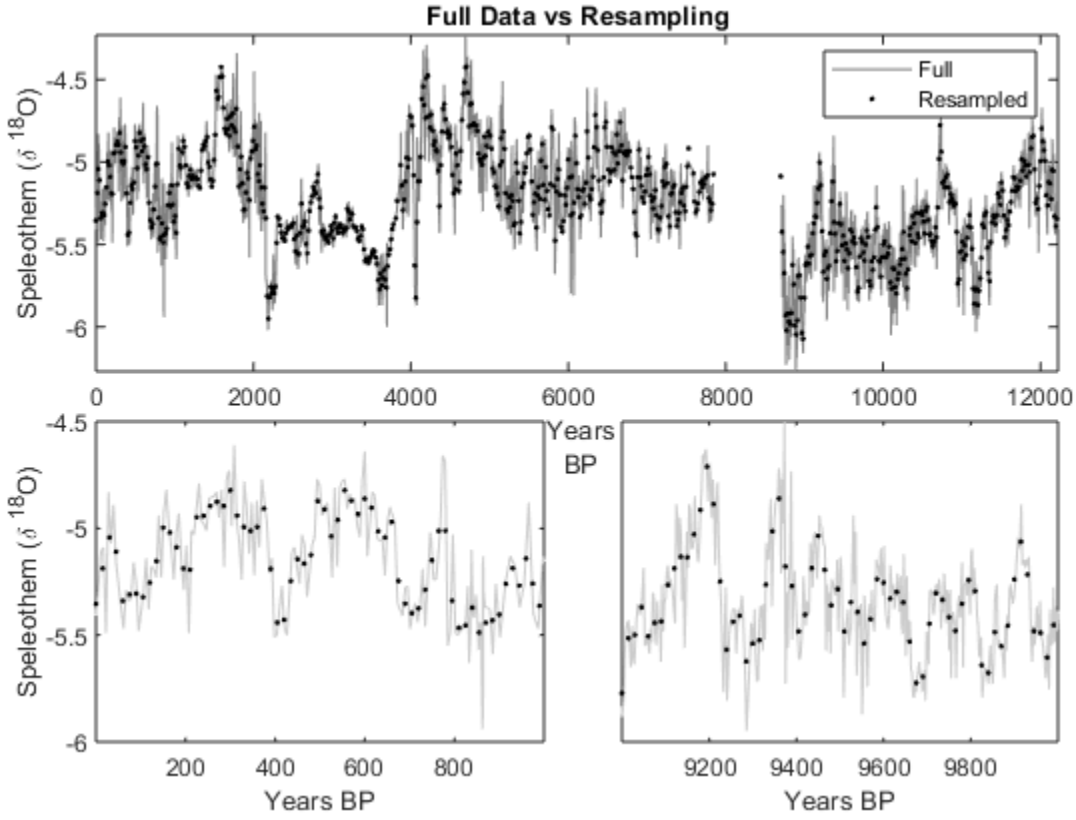

Figure 1. Comparing the full collected data and the resampled data. The two lower plots are zoomed-in to clearly demonstrate the differences between the full data range and the resampled range.

With current DHR, the large gap (872 years) in the data-series needs to be interpolated and due to its size, any interpolation across it is meaningless (in terms of physical interpretation) and could affect the immediate estimates either side of it.

With the proposed ASDHR procedure all that is required is to provide the temporal distance between each pair of samples. This means ASDHR has the full range of data (1919 values) to utilise and ignores the effect of the large gap (there is no interpolation).

3.1.3 Period Identification

The periodicities of the data were identified by scanning across a range of periods to find the two that fit the data best (as in Smith et al., 2016), although as noted above, the current DHR method only uses 815 resampled values while ASDHR has the original 1919 values to use; current DHR method (1290 and 1490 years), ASDHR (1320 and 1540 years), well within the accuracy of the age estimate based on ^{18}O isotope levels. This new result confirms the findings of Smith et al. (2016), only providing a small adjustment when compared to the samples' timing error.

3.1.4 Comparing Current DHR with Proposed ASDHR

To compare the model fit of the two methods, the fit from DHR was rescaled back to the original time base and this showed that the arbitrary sampling procedure yielded slightly better results (Figure 2). The estimated trends and amplitudes were similar between the two methods (Figures 3 and 4 respectively) with the main difference with the behaviour over the gaps. The grey shaded area highlights the large gap in the data and a period of suspect data immediately before it.

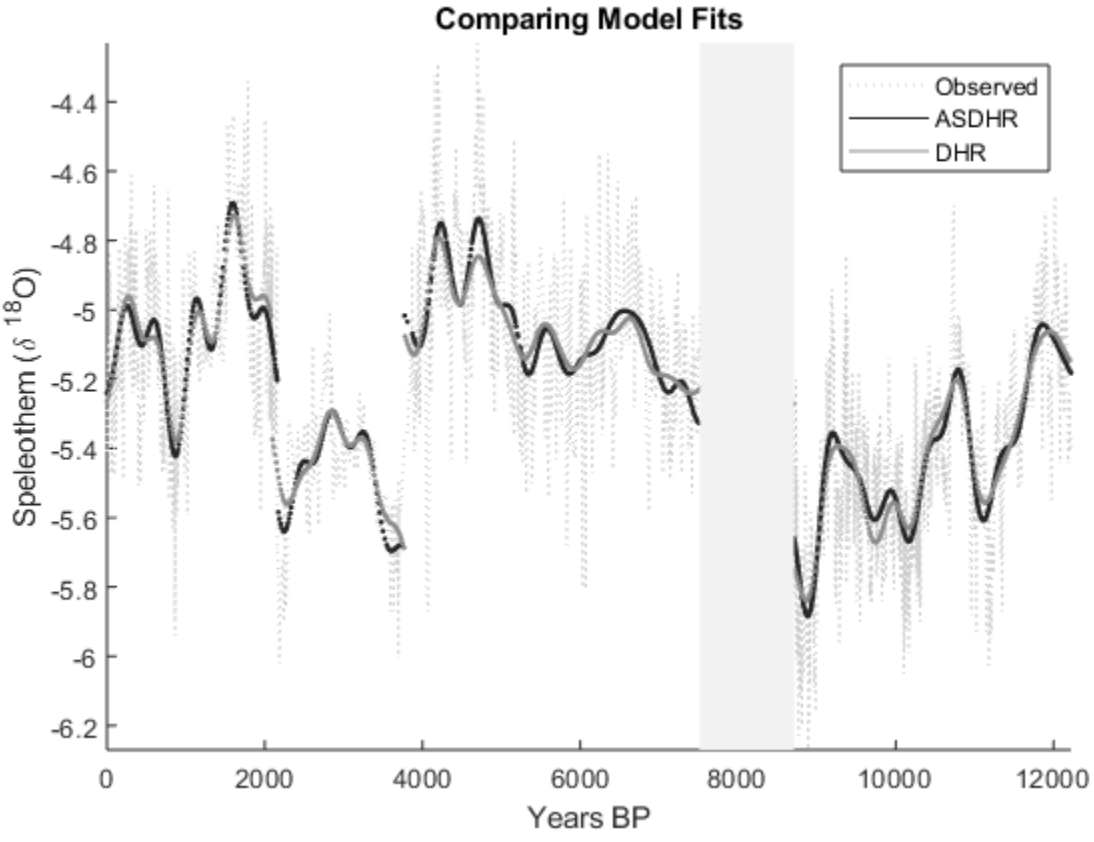

Figure 2. Model fit - comparing ASDHR (R^2 of 0.6741) with DHR (R^2 of 0.6452) on the original time base.

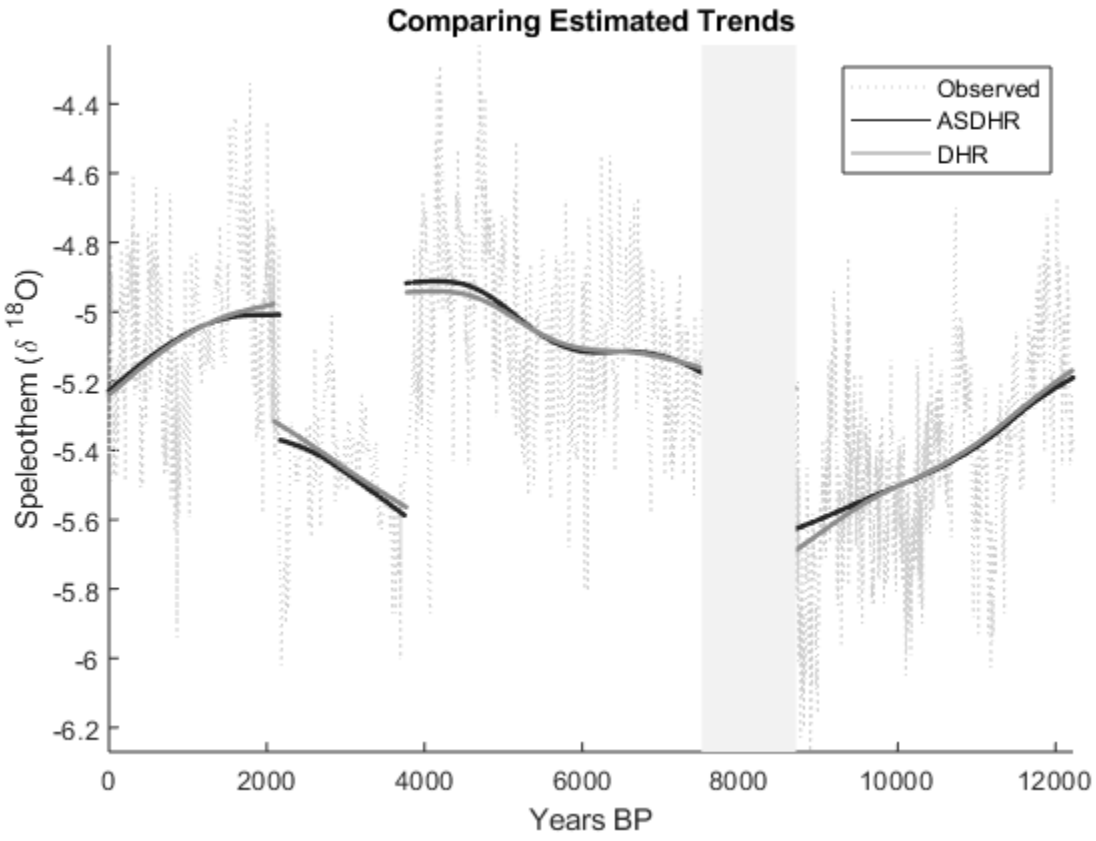

Figure 3. Trend - comparing ASDHR with DHR on the original time base.

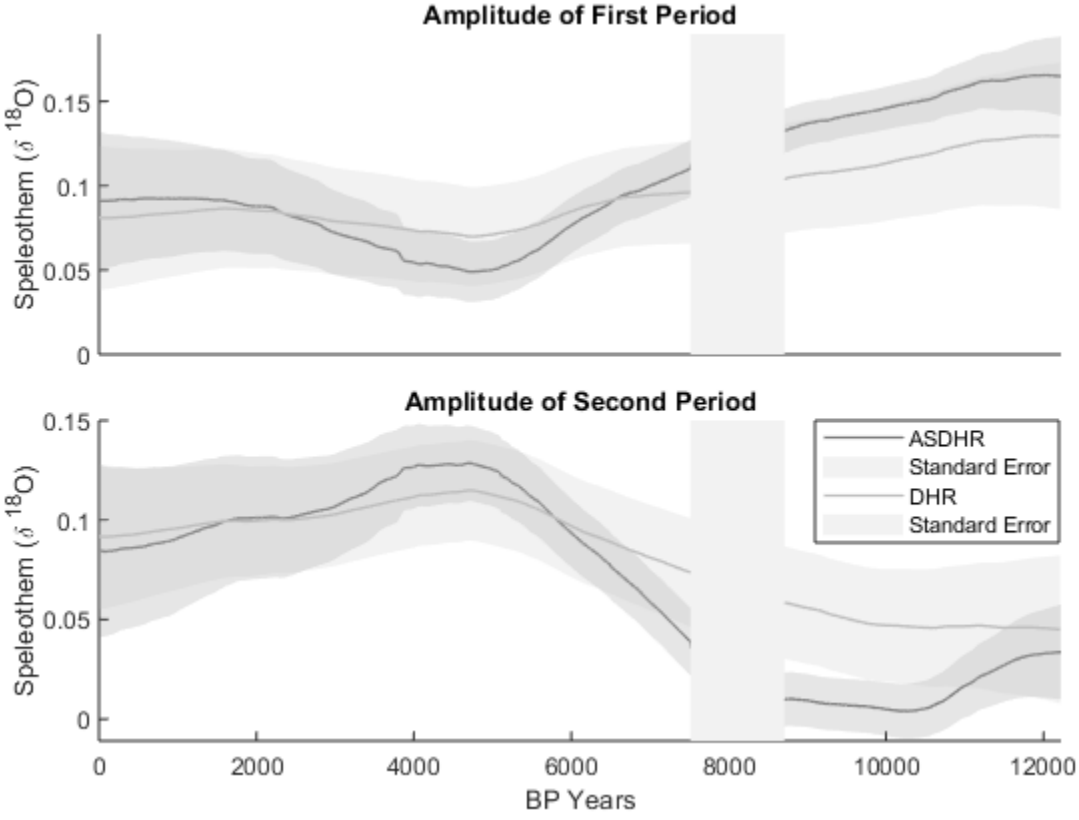

Figure 4. Amplitudes of seasonal components – comparing ASDHR with DHR on the original time base.

The distributions of residual errors were also similar between the two methods, with both being approximately Gaussian and very close to symmetric (Figure 5). This demonstrates how the introduction of the arbitrary sampling technique does not affect the fundamentals of the DHR method.

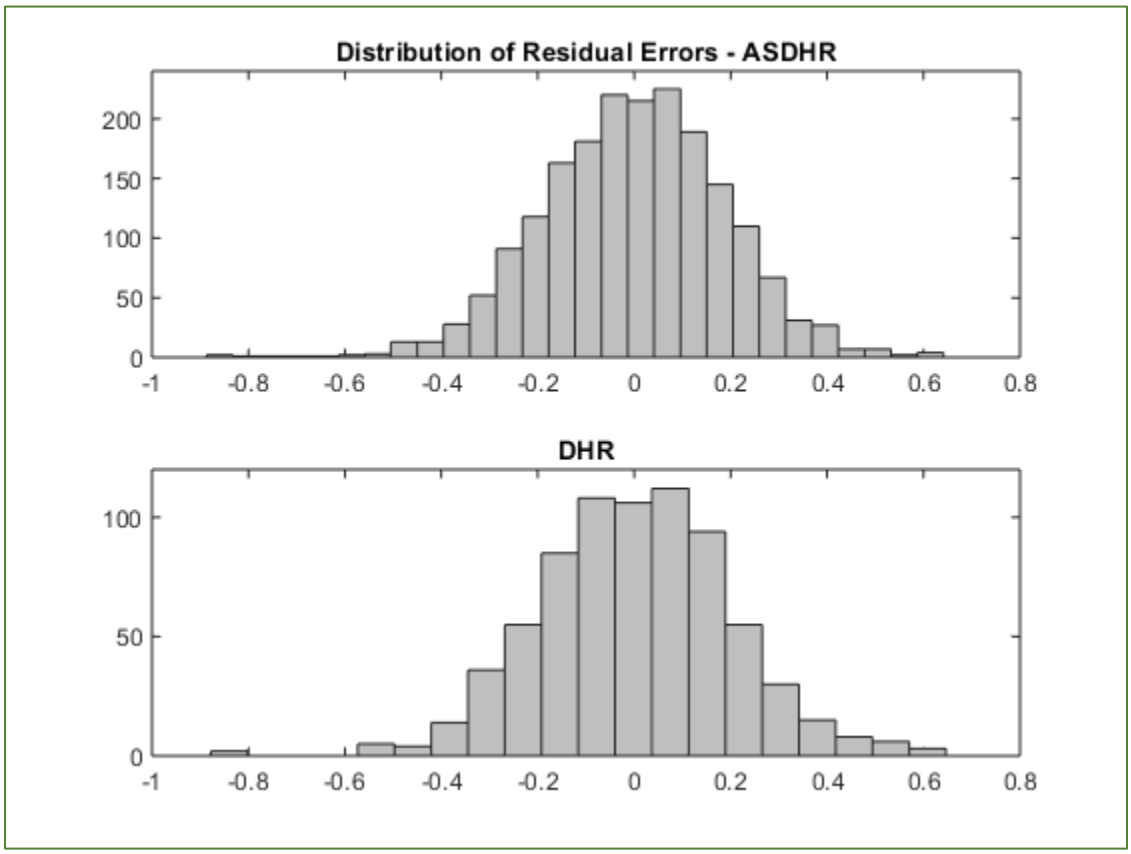

Figure 5. Residual error distributions of both methods.

Using ASDHR for paleo-climatic series not only simplifies the analysis procedure (no prior data manipulation) and preserves the data (i.e. no data loss/artefacts introduced), but also returns slightly better models (in this case a more pronounced seasonal component).

3.2 Persistent Organic Pollutant example

Identifying patterns and trends in atmospheric concentrations of Persistent Organic Pollutants (POPs) is an important part of monitoring and understanding how anthropogenic activities affect them. Weekly air samples have been collected since January 1992 at the High Arctic station of Alert in Canada and are filtered for various POPs and have a high signal to noise ratio (3:1). For further background and DHR analysis see Becker et al. (2006).

Two examples are taken from the data collected at Alert, Benzo(a)pyrene (reported in Becker et al., 2006) and α-Hexachlorocyclohexane (α-HCH, not reported) and both are members of the polycyclic aromatic hydrocarbons subset of POPs. Prior to DHR analysis the data were pre-processed; missing values due to data below the instrumental detection limits were replaced by values 2/3 of the detection limits, and data points situated outside 3x the standard deviation of any fitted trend where considered outliers and removed from the data set. The data were then resampled to fortnightly,

due to the weekly data having significant irregularity, and finally ran through a low pass filter to reduce aliasing.

In the first example the observed data, with missing values and outliers, were used with ASDHR and compared to pre-processed data with DHR. In the second example both use the pre-processed data but without the resampling or pre-filtering for ASDHR. Both model fit and trend were compared between the two techniques as that was the aim of DHR in the original paper.

3.2.1 Benzo(a)pyrene

In Becker et al. (2006) an annual cycle was identified, and using the highly irregularly sampled raw data, this same annual cycle was identified using the ASDHR identification procedure.

The subsequent estimated fit (Figure 6) and trend (Figure 7) show that data pre-processing (unless required for a specific analysis question) is no longer necessary for DHR analysis if the arbitrary sampling technique is used, where a good model fit and trend estimate were obtained from the raw, unfiltered data.

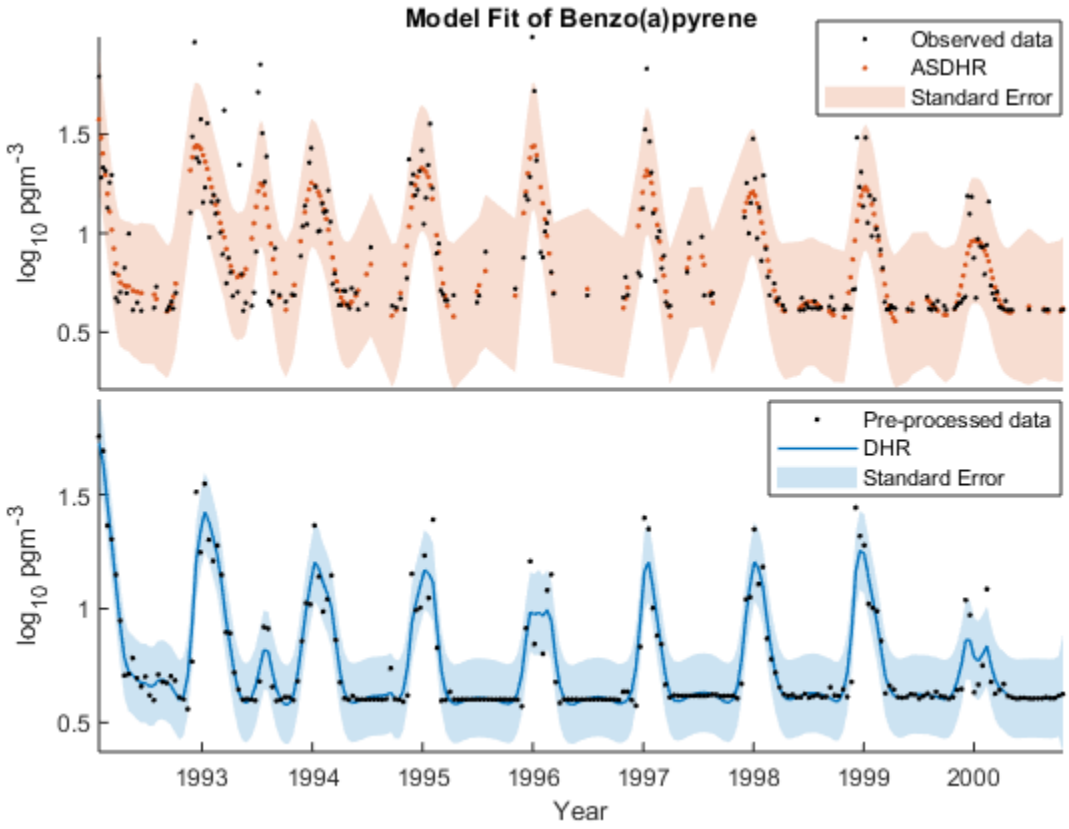

Figure 6. Comparing model fit and data used.

The uncertainty estimate is higher for ASDHR, which can be attributed mainly to the prefiltering used necessarily prior to DHR application, where data were brought onto fortnightly time base. This pre-filtering reduced the variance of the irregular component, which is clearly visible in Figure 6.

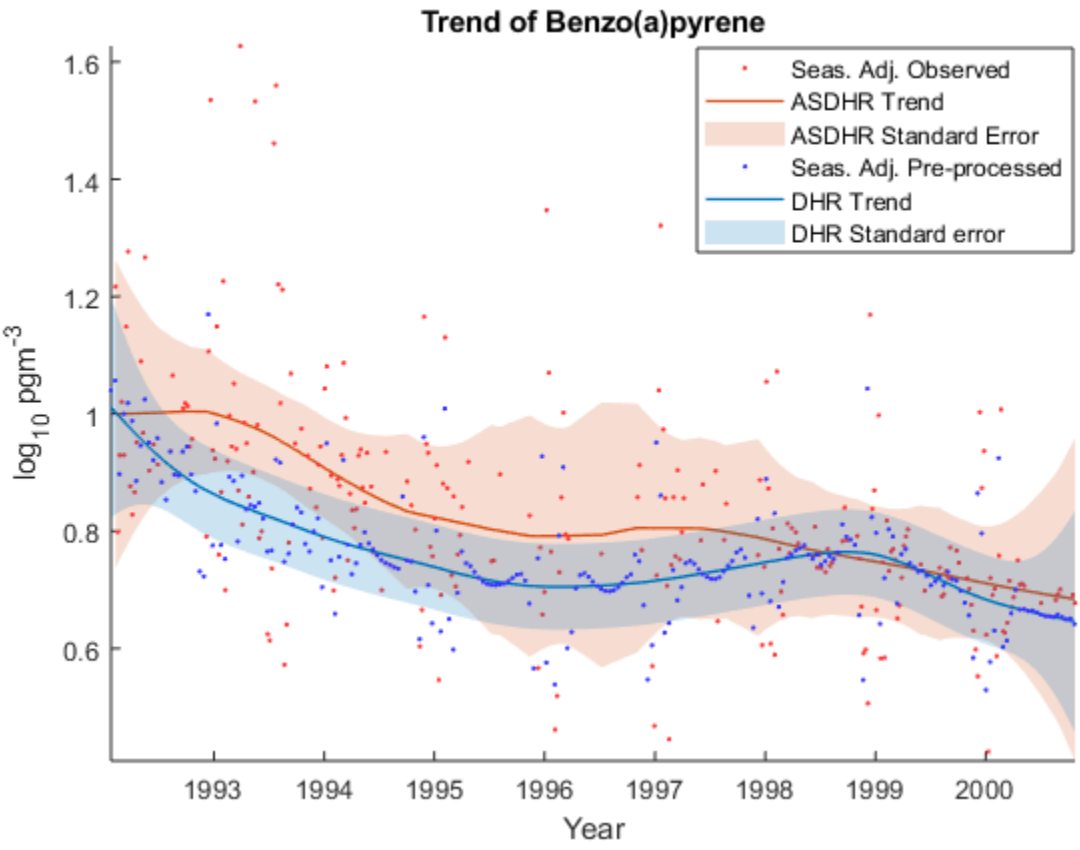

Figure 7. Comparing trend and seasonal adjusted data.

In addition, in ASDHR, as expected, uncertainty grows when data are absent, so the less frequent sampling between 1995 and 1999, as visible in Figure 7, leads to the increase of estimated uncertainty.

3.2.2 α-Hexachlorocyclohexane

Here both DHR and ASDHR used the pre-processed data, where values under detection limits were set to 2/3 of the instrumental detection limits, but for DHR the data were then resampled and pre-filtered.

An annual cycle was identified again, and the subsequent estimation of trend shows how pre-filtering the data can lead to bias in its analysis. In this case (Figure 8), the trend estimated by ASDHR is 'pulled' down by the observed data, data that are missing in the resampled and pre-filtered time-series

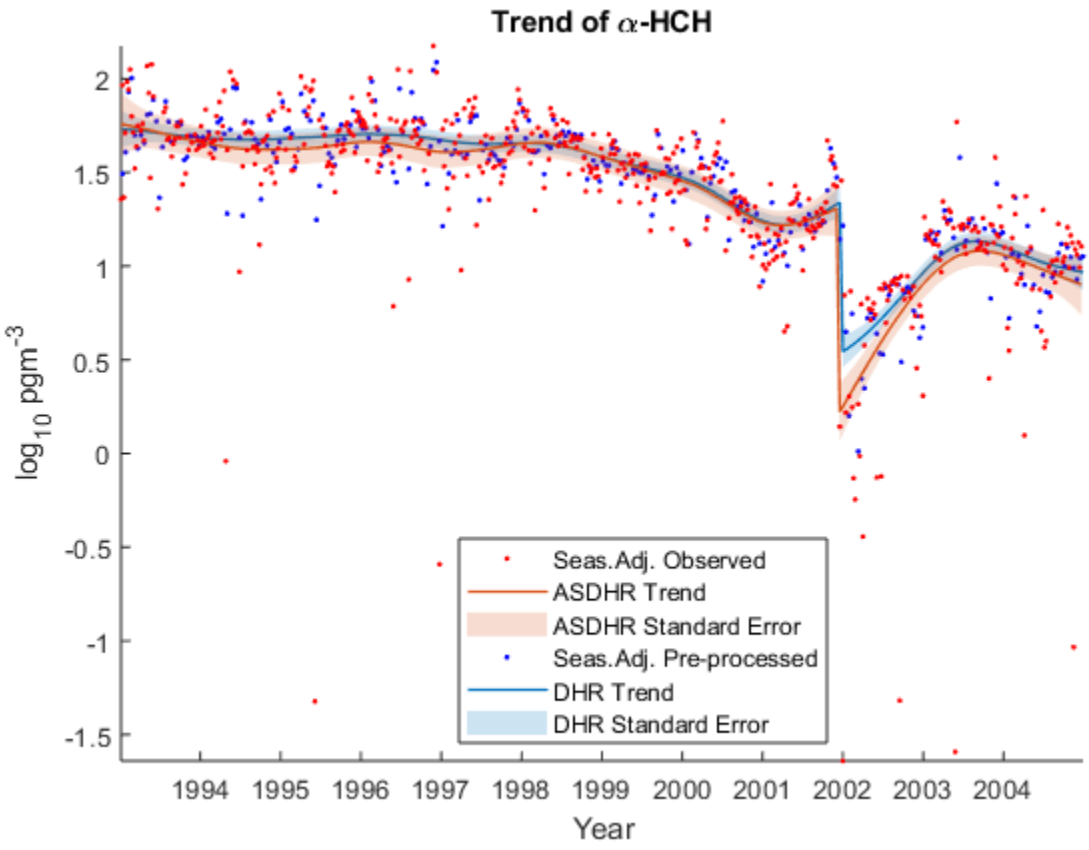

Figure 8. Showing the extent of the data loss from resampling and pre-filtering the data.

3.3 Arbitrary Forecasting – Atmospheric CO_2 example

Below we present a simple example of forecasting of the well-known Keeling Mauna Loa CO2 series (see Acknowledgments). In this example we do not aim at perfect forecasts, but rather at showing the flexibility and versatility of even the basic version ASDHR in this application. Better forecasts could be achieved with more assumptions being included in the model, such as a model of the business cycle or the industrial growth projections.

The Mauna Loa CO_2 monthly mean data set was used to demonstrate arbitrary forecasting, that is forecasting from any point in the future and not just from the end of the data-set. Additionally, the time-base of the forecasting period is not limited to that of the observed data, something that the original DHR cannot do.

This is easily implemented by extending vector Δ_k by the values corresponding to the arbitrary points chosen. In this example (Figure 9) the monthly Mauna Loa data are used to forecast the year 2018 on a weekly basis, for five different periods of data: 1960 to 1970, to 1980, to 1990, to 2000, and to 2010 as shown with colour code in Figure 8. In this example the forecast shown as purple is based solely on the data up to 1970.

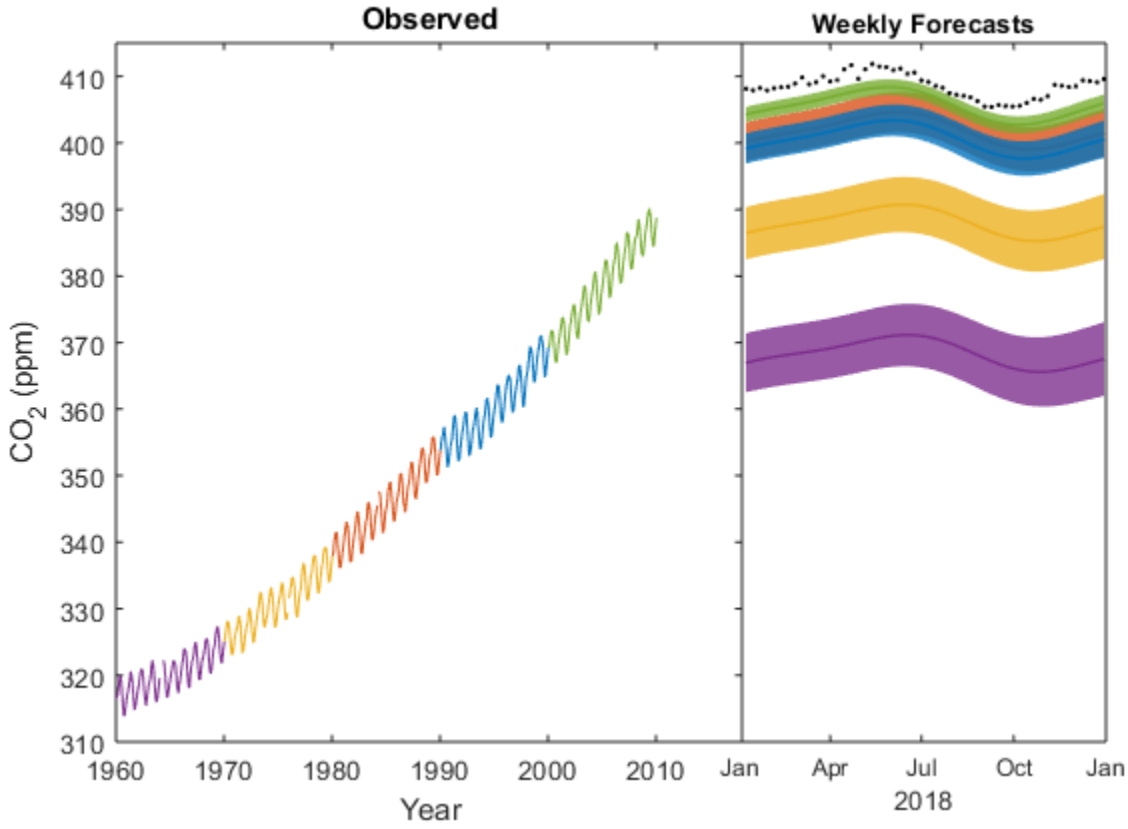

Figure 9. Arbitrary forecasting of Mauna Loa CO_2 data. Observed data is monthly averages and forecasts are weekly averages. Purple, 1960:1970. Yellow, 1960:1980. Orange, 1960:1990. Blue, 1960:2000. Green, 1960:2010. Shading, uncertainty of each forecast. Black dots, observed weekly average data for 2018.

This (Figure 9) demonstrates how the rate of change of CO_2 grows with time – as the knowledge is increasing (e.g. the 1960:1970 based predictions for 2018 have levels observed in 2000), except during the 1990s where the 1960:2000 based predictions are lower than that of the 1960:1990, clearly relating to the visible dip during the 1990s.

The size of the uncertainty, shown here as a single standard error estimate band, increases as the forecasting horizon increases; purple for 48 years, green for 8 years. Testing on this data set and on simulated data sets indicates that the forecasts are largely unaffected by data sparseness or the forecasting horizon magnitude. Only the observation noise level impacts the predictions.

In the past the DHR technique was successfully used for forecasting various processes, in both broadly environmental and industrial applications: from demand for electricity (Young and Pedregal, 1998) to numbers of calls to a banking call centre (Tych et al, 2002) to, anecdotally, demand for beer (unpublished Lancaster University thesis). ASDHR naturally broadens the application area of this robust method.

4.0 Robustness and reliability evaluation

Before the methodology was applied to real data, it was tested on a variety of challenging simulated data-sets (with known “unobserved” components) to test the robustness and sensitivity to data sparseness, size of temporal distancing, and observational noise.

The methodology was found to be very robust with very sparse data-sets. It also produces meaningful estimates for a very wide range of temporal distances between the samples, limited by the Taylor expansion (3) and condition of the recursive estimate of the covariance matrix within the Fixed Interval Smoother algorithm applied. Subject to these constraints, the technique can be used to estimate the observed values and components no matter how irregularly sampled the time-series is.

Sensitivity to observation noise is similar between the time-tested and commonly used DHR (with about 300 citations) and ASDHR.

In terms of physical interpretation, the extent of data sparseness needs to be considered when interpreting the estimated components as very sparse data may effectively under-sample higher frequency components.

5.0 Conclusions

Presented here is a technique to improve the DHR analysis of irregularly sampled time-series that removes the need for data pre-processing: regularisation, decimation, interpolation etc. The technique also does not involve estimation of missing values and so any subsequent analysis is free from assumptions and bias and only uses the available observed data. This brings DHR closer to the DBM philosophy of allowing the data to inform us of the processes and mechanisms that result in the observed time-series. It also makes it uniquely suitable for analysis of environmental irregularly sampled observational data.

Data pre-processing was a necessary step to allow DHR to work on irregularly sampled data-sets but it comes with artefacts, bias and increased uncertainties in the model estimates. However, with the arbitrary sampling technique, this step can be avoided and provides model estimates with lower uncertainties and no bias.

The technique has been tested on challenging simulated data and is robust enough to work on extremely sparse data, however, in terms of physical interpretation there is a limit to how sparse the data can be due to e.g., under-sampling (Chappell et al., 2017). Additionally, the technique was found to have similar observation noise sensitivity to that found in standard DHR method.

The technique has been demonstrated here on three different types of observed environmental time-series data and has yielded slightly better model outputs than the standard DHR method. Without data pre-processing, there will be no introduction of any assumptions, artefacts or bias into the data prior to analysis and thus these results should be closer to observed reality.

The technique also allows for forecasting at arbitrary points and at different sampling rates than in the observed data. This means the frequency of the forecast is not limited to the frequency of the observations, and with a non-stationary forecast horizon may allow forecasting to yield more insights into environmental processes.

Finally, while it may not be apparent from the equations, ASDHR is easily and inherently generalised so that all aspects of estimation are either time-varying or state-dependent: from periodicity to dynamics of random walk model, to NVRs. For each sample k, the periodicity, random walk model and NVR can be set. So, for example, one section of observed data could be analysed for one set of periodicities and another section analysed for another set, or the random walk model could be changed to match a significant change in the data, or the NVRs can be varied to suit the smoothness of the data.

Software format and availability

The method has been implemented within the Matlab computing environment. The input arguments specify the time-series variable, the State-Space format, meta-parameters for the trend and for harmonics' amplitudes, which essentially specify the time-scale of the trend and amplitude variability. Additional arguments control variance interventions, initial conditions etc, with sensible default values provided. The output arguments have also been constructed with ease of use in mind, and include model fit, estimated trend, harmonic components and their amplitudes with their respective uncertainty estimates, so are immediately interpretable in terms of the modelled environmental process.

Matlab functions and example scripts are available from the corresponding author upon request. The functions will be included in the CAPTAIN Toolbox for Matlab in due course.

Acknowledgements

Paleo-climatic data (Smith et al., 2016) have been kindly provided by Dr Andi Smith of British Geological Survey.

The authors acknowledge Dr Hayley Hung, Environment Canada, for access to the Northern Contaminants Program (NCP) and the Integrated Atmospheric & Deposition Network (IADN) datasets for persistent organic pollutants (POPs).

Dr. Pieter Tans, NOAA/ESRL (www.esrl.noaa.gov/gmd/ccgg/trends/) and Dr. Ralph Keeling, Scripps Institution of Oceanography (scrippsco2.ucsd.edu/).

References

Becker, S., Halsall, C. J., Tych, W., Hung, H., Attewell, S., Blanchard, P., Li, H., Fellin, P., Stern, G., Billeck, B., Friesen, S. 2006. Resolving the Long-Term Trends of Polycyclic Aromatic Hydrocarbons in the Canadian Arctic Atmosphere. *Environmental Science & Technology*, 40, 3217-3222.

Bhar, R. 2010, *Stochastic Filtering with Applications in Finance*, World Scientific Publishing

Box, G. E. P., Tiao, G. C. 1975. Intervention Analysis with Applications to Economic and Environmental Problems. *Journal of the American Statistical Association*, 70 (349), 70-79.

Brockwell, P. J. (2001). Levy-Driven Carma Processes. *Ann. Inst. Statist. Math*, 53 (1), 113-124.

Broersen, P. M. T., de Waele, S., Bos, R. (2004). Autoregressive spectral analysis when observations are missing. ***Automatica***, 40 (9), 1495-1504.

Carling, P.A., Tych, W., Richardson, K. 2005, The hydraulic scaling of step-pool systems, *River, coastal and estuarine morphodynamics, in Parker, G. and Garcia, M.H. (eds.), 144, 55-63, Taylor and Francis*

Chappell, N. A., Tych, W. 2012. Identifying step changes in single streamflow and evaporation records due to forest cover change. *Hydrological Processes*, 26 (1), 100-116.

Chappell, N.A., Jones, T.D., Tych, W. 2017. Sampling frequency for water quality variables in streams: systems analysis to quantify minimum monitoring rates. *Water Research*, 123, 49-57.

Fan, J., Gijbels, I. 1996. Local polynomial modelling and its applications: monographs on statistics and applied probability 66, (Vol. 66). CRC Press.

Fisher, R. A. 1929. Tests of Significance in Harmonic Analysis. *Proceedings of the Royal Society of London, A,* 125, 54-59.

Halliday, S. J., Skeffington, R. A., Wade, A. J., Neal, C., Reynolds, B., Norris, D., Kirchner, J. W. 2013. Upland streamwater nitrate dynamics across decadal to sub-daily timescales: a case study of Plynlimon, Wales. *Biogeosciences*, 10, 8013-8038.

Keery, J., Binley, A., Crook, N., Smith, J. W.N. 2007. Temporal and spatial variability of groundwater-surface water fluxes: Development and application of an analytical method using temperature time series. *Journal of Hydrology*, 336 (1-2), 1-16.

Li, W., Shah, S.L., 2008, Kalman Filters in non-uniformly sampled multirate systems, *Automatica,* 44, 199-208

Mathias, A., Grond, F., Guardans, R., Seese, D., Canela, M., Diebner, H.H. 2004, Algorithms for spectral analysis of irregularly sampled time series, *Journal of Statistical Software*, 11 (2),

Mindham, D. A., Tych, W., Chappell, N. A. 2018. Extended State Dependent Parameter modelling with a Data-Based Mechanistic approach to nonlinear model structure identification. *Environmental Modelling & Software*, 104 (2018), 81-93.

O'Toole, S. J., Butler, R. P., Tinney, C. G., Jones, H. R. A., Marcy, G. W., Carter, B., McCarthy, C., Bailey, J., Penny, A. J., Apps, K., Fischer, D. 2007. New Planets around Three G Dwarfs. *Astrophysics Journal*, 660 (2), 1636-1641.

Smith, A, C., Wynn, P. M., Barker, P. A, Leng, M. J., Noble, S. R., Tych, W. 2016. North Atlantic forcing of moisture delivery to Europe throughout the Holocene. *Scientific Reports*, 6:24745.

Sofianopoulou, E., Pless-Mulloli, T., Rushton, S, Diggle, P. J. 2017, Modelling seasonal and spatiotemporal variation: the example of respiratory prescribing, *American Journal of Epidemiology*, 186 (1), 101-108.

Taylor, C. J., Pedregal, D. J., Young, P. C., Tych, W. 2007. Environmental time series analysis and forecasting with the CAPTAIN toolbox. *Environmental Modelling & Software*, 22, 797–814.

Trapero, J. R., Kourentzes, N., Martin, A. 2015. Short-term Solar Irradiation forecasting based on Dynamic Harmonic Regression. *Energy*, 84, 289-295.

Tych, W., Pedregal, D.J., Young, P.C., Davies, J. 2002, An unobserved component model for multi-rate forecasting of telephone call demand: the design of a forecasting support system, *International Journal of Forecasting*, 18, 673-695.

UNEP, 2011. Climate Change and POPs: Predicting the Impact. Report of the UNEP/AMAP Expert group, Geneva, Switzerland.

Venier, M., Hung, H., Tych, W., Hites, R. A. 2012. Temporal Trends of Persistent Organic Pollutants: A Comparison of Different Time Series Models. *Environmental Science & Technology*, 46 (7), 3928-3934.

Young, P. C., Ng, C. 1989. Variance intervention. *Journal of Forecasting*, 8 (4), 399-416.

Young, P.C., Pedregal, D.J. 1998. Adaptive Electricity Demand Forecasting Using a Novel Non-Linear, Unobserved Components Model Estimated in the Frequency Domain, *Colloquium on Electricity Demand Forecasting*, London Business School, London, July 6th 1998.

Young, P. C., Pedregal, D. J., Tych, W. 1999. Dynamic Harmonic Regression. *Journal of Forecasting*, 18, 369-394.

Young, P. C. 1999b. Data-based mechanistic modelling, generalised sensitivity and dominant model analysis. *Computer Physics Communications*, 117, 113-129.