Ross Whiteford[1], Timothy J. Heaton[2], Michael J. Henehan[3], Eleni Anagnostou[4], Hana Jurikova[1], Gavin L. Foster[5], and James W.B. Rae[1]

[1]University of St Andrews

[2]Department of Statistics, School of Mathematics, University of Leeds, Leeds LS2 9JT, UK

[3]School of Earth Sciences, University of Bristol, Wills Memorial Building, Queens Road, Bristol, UK

[4]GEOMAR Helmholtz Centre for Ocean Research Kiel

[5]School of Ocean Earth Science, University of Southampton, UK

Contents of this file

1. Text S1
2. Figures S1 to S4
3. Data files S1 to S3

Additional Supporting Information (Files uploaded separately)

1. Data file S1 - Reconstructed $\delta^{11}B_{sw}$ and summary metrics
2. Data file S2 - Reconstructed pH summary metrics
3. Data file S3 - Data and reconstructed $^{87/86}Sr$, $\delta^{7}Li$, and $^{187/188}Os$.

1. Gaussian Processes With Non-Gaussian Constraint Noise

1.1. Introduction

In the main text, we gave a brief overview of the Gaussian Process methodology as it related to reconstruction of $\delta^{11}B_{sw}$. Here we give a fuller, more generalised and statistically rigorous description of the Gaussian Process methodology, including how it was adapted to incorporate each style of constraint mentioned in the main text. We then illustrate how this technique works in practice by testing it against a hypothetical signal with noisy constraints, with a step by step walkthrough of integrating various forms of information analogous to the types of constraint we have on $\delta^{11}B_{sw}$.

Suppose that we observe a function $f(\cdot)$ subject to (potentially non-Gaussian) noise in constraints at a set of known times t_i, i.e.,

$$y_i = f(t_i) + \epsilon_i \quad \text{for } i = 1, \ldots, M.$$

Here, ϵ_i can be a general probability distribution and is not required to be a standard Gaussian. Furthermore, suppose that we may have some further constraints on the value of $f(t)$ for certain values of t (e.g. lower/upper bounds), or additional non-standard information on the values of $f(t)$ (such as constraints on the change in value over time $\frac{df}{dt}$).

We wish to obtain a non-parametric posterior estimate of the function $f(\cdot)$ modelled as a Gaussian Process (GP) given both the potentially complex observations y_i and any additional non-standard information. In the standard GP setting, the function $f(\cdot)$ is assumed to be observed subject to normally distributed noise. As a consequence, the exact posterior for f given $\mathbf{y}$ can be easily calculated directly. However, when the available

constraints ϵ are non-Gaussian, the GP's posterior can not be written down exactly and is much more challenging to calculate, as it does not take a standard form. To obtain posterior samples under such a non-Gaussian constraint model, we therefore implement a rejection sampling approach. Specifically, we aim to draw from a *nearby* distribution (from which it is possible to sample directly) and then reject/accept these samples using rejection sampling principles to obtain the correct GP posterior under our non-Gaussian constraint model.

To explain our approach in this Supplementary Information, we will first provide a brief background to Gaussian Processes and explain how they are usually fitted in the context of non-parametric regression given a set of constraints with Gaussian noise. We then introduce the idea of rejection sampling before going on to show how this idea can be used to sample from the posterior of a Gaussian Process in the presence of constraints with non-Gaussian uncertainty (or when additional, non-standard information is available).

1.2. Definition of a Gaussian Process (GP) Prior

A (one-dimensional input) zero-mean *Gaussian Process* $f(z) \sim \mathcal{GP}(0, k(t,t'))$ is a collection of random variables, any finite number of which have a joint Gaussian distribution (Rasmussen & Williams, 2006). It is completely specified by its covariance function:

$$k(t,t') = \mathbf{E}[f(t)f(t')].$$

When using a Gaussian Process to perform regression, the random variables represent the values of the function $f(t)$ at time t. For a set of N times $\mathbf{t}_\star = (t_1^\star, t_2^\star, \ldots, t_N^\star)^T$, our prior specifies

$$\mathbf{f}_\star = \mathbf{f}(\mathbf{t}_\star) \sim \mathcal{N}(0, K_{\mathbf{t}_\star,\mathbf{t}_\star}),$$

where $K_{\mathbf{t}_\star,\mathbf{t}_\star}$ denotes the matrix of the covariances evaluated at all pairs of the times $t_i^\star$.

1.3. Updating the GP Prior under a Normal Observational Model

Typically, when performing non-parametric regression, we assume that we observe the function $f(t)$ subject to normally-distributed noise, i.e.,

$$\mathbf{y} = \mathbf{f}(\mathbf{t}) + \boldsymbol{\eta}$$

where the noise $\boldsymbol{\eta} \sim \mathcal{N}(0, \Sigma)$. This observational noise may have dependence encoded in the covariance matrix Σ but critically is assumed to be normally distributed. In such a situation, we can use the standard properties of the multivariate normal distribution to derive the posterior distribution for our function values at our times of interest $\mathbf{t}_\star$ exactly:

$$\mathbf{f}_\star|\mathbf{t}, \mathbf{y}, \mathbf{t}_\star \sim \mathcal{N}(\bar{\mathbf{f}}_\star, cov(\mathbf{f}_\star)),$$

where

$$\bar{\mathbf{f}}_\star = K_{\mathbf{t},\mathbf{t}_\star}^T [K_{\mathbf{t},\mathbf{t}} + \Sigma]^{-1} \mathbf{y},$$

$$cov(\mathbf{f}_\star) = K_{\mathbf{t}_\star,\mathbf{t}_\star} - K_{\mathbf{t},\mathbf{t}_\star}^T [K_{\mathbf{t},\mathbf{t}} + \Sigma]^{-1} K_{\mathbf{t},\mathbf{t}_\star}.$$

See Rasmussen and Williams (2006) for full details. However, as soon as the observational model becomes non-normal, i.e., it is no longer the case that $y_i|f(t_i) \sim N(f(t_i), \sigma_i^2)$ then the GP posterior becomes much more complex and will no longer take the form of a simple multivariate normal. To estimate the posterior distribution in such instances, we will therefore take a different approach based upon rejection sampling.

1.4. Rejection Sampling

Rejection sampling is a general purpose method that enables sampling from non-standard distributions. Suppose we wish to sample $\mathbf{X}$ from a particular target probability

density $f_X(\mathbf{x})$ but, for some reason, we cannot do so directly. However, suppose there exists an alternative *envelope* density function $g_Z(\mathbf{z})$ from which we can sample that satisfies the condition $\frac{f_X(\mathbf{x})}{g_Z(\mathbf{x})}$ bounded $\forall \mathbf{x}$. For any constant $c \geq \sup_{\mathbf{x}} \frac{f_X(\mathbf{x})}{g_Y(\mathbf{x})}$, we can then obtain samples from our desired target $f_X(\mathbf{x})$ using the following rejection method:

1. Sample $\mathbf{z}$ from an envelope density that is proportional to $g_Z(\mathbf{z})$, and a uniform u from $U[0, 1]$.

2. If $u \leq \frac{f_X(\mathbf{z})}{c\, g_Y(\mathbf{z})}$, state $\mathbf{X} = \mathbf{z}$, otherwise return to step 1.

For maximum efficiency we therefore want c, i.e., $\sup_{\mathbf{x}} \frac{f_X(\mathbf{x})}{g_Z(\mathbf{x})}$ as small as possible. We therefore aim to find an *envelope* density function g_Z that is both easy to sample from and mimics the target f_X as closely as possible. The concept of rejection sampling is shown graphically in Figure S3.

2. Rejection Sampling GPs with Non-Gaussian Noise

Returning to our specific non-parametric regression, suppose that we observe a function subject to (potentially non-Gaussian) noise at a set of known times t_i,

$$y_i = f(t_i) + \epsilon_i \quad \text{for } i = 1, \ldots, M.$$

We wish to place a Gaussian Process prior on the values of $f(t)$ and then sample from the posterior under the (potentially non-Gaussian) observational model ϵ,

$$p_\epsilon(\mathbf{f}|\mathbf{y}) = \frac{p_\epsilon(\mathbf{y}|\mathbf{f})\pi(\mathbf{f})}{p_\epsilon(\mathbf{y})} \propto p_\epsilon(\mathbf{y}|\mathbf{f})\pi(\mathbf{f}). \qquad (\dagger)$$

Due to the non-Gaussian nature of ϵ, we cannot directly sample from this posterior distribution. However we can sample from an alternative distribution of our choosing and then use rejection sampling principles. We will typically use the GP posterior under a

normally-distributed error model for this envelope distribution. This η-error model as discussed in Section 1.3 can be calculated precisely and, in general, will hopefully be close to the true target posterior. Having chosen a suitable envelope density, proportional to $g(\mathbf{f}|\mathbf{y})$, the rejection algorithm thus becomes:

1. Calculate $c^{\ddagger} = \sup_{\mathbf{f}} \frac{p_{\epsilon}(\mathbf{y}|\mathbf{f})\pi(\mathbf{f})}{g(\mathbf{f}|\mathbf{y})}$

2. Sample from the envelope density $g(\mathbf{f}^{\ddagger}|\mathbf{y})$ a potential $\mathbf{f}^{\ddagger}$ at both times of interest $\mathbf{t}^{\star}$ and the times $\mathbf{t}$ at which we have observations $\mathbf{y}$

3. Sample $u \sim U[0,1]$, if $u \leq \frac{p_{\epsilon}(\mathbf{y}|\mathbf{f}^{\ddagger})\pi(\mathbf{f}^{\ddagger})}{c^{\ddagger}g(\mathbf{f}^{\ddagger}|\mathbf{y})}$ then accept $\mathbf{f} = \mathbf{f}^{\ddagger}$ as a draw from the correct posterior, otherwise return to step 2.

The calculation of both $c^{\ddagger}$ and the acceptance criteria in step 3 will generally only depend upon the sampled values of $\mathbf{f}^{\ddagger}$ at the times $\mathbf{t}$ with observations. This will reduce calculation. Furthermore, we note that $p_{\eta}(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{N} p_{\eta_i}(y_i|f(t_i))$ if the observations are independent. Exceptions may however occur if we have additional, non-standard, constraints such as on the range or variation of the function $f(\cdot)$.

We can repeat this rejection sampling technique until we obtain a large number of posterior realisations $\mathbf{f}$ from the target distribution (that corresponds to the general observational noise model). These can then be summarised by Monte Carlo to provide posterior means and variances for any $f(t)$.

2.1. Rejection Sampling Implementation

Our rejection sampling algorithm to sample from the correct posterior under a general observational error model ϵ then becomes (after cancelling common terms):

1. Sample $\mathbf{f}^\eta$ at times of interest $\mathbf{t}^\star$ and also times $\mathbf{t}$ at which we have observations $\mathbf{y}$ from GP posterior under normally-distributed η error model.

2. Sample $u \sim U[0,1]$, if $u \leq \frac{p_\epsilon(\mathbf{y}|\mathbf{f}^\eta)}{c^\star\, p_\eta(\mathbf{y}|\mathbf{f})}$ (where, as defined above, $c^\star = \sup_{\mathbf{f}} \frac{p_\epsilon(\mathbf{y}|\mathbf{f}^\eta)}{p_\eta(\mathbf{y}|\mathbf{f})}$) then accept $\mathbf{f} = \mathbf{f}^\eta$ as a draw from the correct posterior, otherwise return to step 1. Again, we note that this only depends upon the sampled values of $\mathbf{f}^\eta$ at the times $\mathbf{t}$ with observations and that, e.g., $p_\eta(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^N p_{\eta_i}(y_i|f(t_i))$ if those observations are independent.

We repeat this sampling technique until we obtain a large number of posterior realisations $\mathbf{f}$ from the target distribution (that correspond to the general observational noise model) which can then be summarised by Monte-Carlo.

3. Specific Examples

While the rejection sampling approach may appear complicated, in many instances it will simplify considerably. We discuss some specific examples below.

3.1. Incorporating Upper and Lower Bounds

Suppose that we have a set of normally-distributed observations $\mathbf{y}$ but, in addition, a further set of values $\mathbf{z} = (z_1, \ldots, z_K)^T$ that operate as upper bounds on the unknown function, i.e., it is the case that $f(t_j^b) < z_j$ for given times $t_1^b, \ldots, t_K^b$. In this case, we consider that these K additional values are entirely uninformative about the value of $f(t)$ beyond providing such a bound. Consequently, the target posterior is:

$$p(\mathbf{f}|\mathbf{y}, \mathbf{z}) \propto \left\{ \prod_{j=1}^{M} \mathbb{1}_{[f(t_j^b)<z_j]} \right\} p_\eta(\mathbf{y}|\mathbf{f})\pi(\mathbf{f}),$$

where $p_\eta(\mathbf{y}|\mathbf{f})$ is the usual normal likelihood function for the observations $\mathbf{y}$. For our envelope function, we can sample directly from the GP posterior considering just the

regular, normally-distributed, observations $\mathbf{y}$, i.e., $g(\mathbf{f}|\mathbf{y}) \propto p_\eta(\mathbf{y}|\mathbf{f})\pi(\mathbf{f})$ so that our $c^\ddagger = 1$. Our algorithm then becomes simply:

1. Sample $\mathbf{f}^\ddagger$ from the standard GP posterior based upon normally-distributed observations $\mathbf{y}$ at both times of interest $\mathbf{t}^\star$ and the times $\mathbf{t}^b$ at which there are upper bounds. This can be done as described in Section 1.3

2. Accept $\mathbf{f}^\ddagger$ as a draw from the true target posterior if it satisfies all the constraints $\mathbf{z}$; otherwise reject and return to step 1.

This has a straightforward analogue when we have combinations of upper and lower bounds.

3.2. Incorporating Non-Gaussian Observations

When our observations $\mathbf{y}$ are subject to non-normal noise (which we have denoted by η) then an appropriate envelope density to use for rejection sampling might be the GP posterior for $\mathbf{f}$ had the noise been normally distributed (see Figure S3). In other words, we use might use a GP conditioned on observations with normal noise as our initial estimate of the posterior, then refine this through rejection sampling, i.e., the posterior for $f(\cdot)$ under the model:

$$y_i = f(t_i) + \eta_i \quad \text{for } i = 1, \ldots, n.$$

where $\eta \sim N(0, \sigma_i^2)$. In this case the envelope function is $g(\mathbf{f}|\mathbf{y}) = p_\eta(\mathbf{f}|\mathbf{y}) = p_\eta(\mathbf{y}|\mathbf{f})\pi(\mathbf{f})$. This distribution is known, see Section 1.3, and it is easy to sample from it directly. To perform rejection sampling, we are required to calculate

$$c^\dagger = \sup_{\mathbf{f}} \frac{p_\epsilon(\mathbf{y}|\mathbf{f})\pi(\mathbf{f})}{p_\eta(\mathbf{y}|\mathbf{f})\pi(\mathbf{f})} = \sup_{\mathbf{f}} \frac{p_\epsilon(\mathbf{y}|\mathbf{f})}{p_\eta(\mathbf{y}|\mathbf{f})}.$$

Since $p_\eta(\mathbf{y}|\mathbf{f})$ is a normal distribution with infinite support, this supremum will exist for almost all alternative errors models (unless they have different tail behaviour). The calculation of $c^\dagger$ only depends upon the sampled values of $\mathbf{f}$ at the times $\mathbf{t}$ for which we have observations $\mathbf{y}$. Furthermore, if the observations y_i are independent, the numerator and denominator in the supremum can be calculated as independent products since, e.g., $p_\eta(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^N p_{\eta_i}(y_i|f(t_i))$. Our rejection sampling algorithm to sample from the correct GP posterior under a general observational error model ϵ then becomes:

1. Sample $\mathbf{f}^\ddagger$ at times of interest $\mathbf{t}^\star$ and also times $\mathbf{t}$ at which we have observations $\mathbf{y}$ from GP posterior under normally-distributed η error model.

2. Sample $u \sim U[0,1]$, if $u \leq \frac{p_\epsilon(\mathbf{y}|\mathbf{f}^\ddagger)}{c^\dagger\, p_\eta(\mathbf{y}|\mathbf{f}^\ddagger)}$, then accept $\mathbf{f} = \mathbf{f}^\ddagger$ as a draw from the correct posterior. Otherwise return to step 1.

Again, the acceptance criteria in step 2 only depends upon the sampled values of $\mathbf{f}^\ddagger$ at the times $\mathbf{t}$ corresponding to the observations $\mathbf{y}$. Also, if the observations $\mathbf{y}$ are independent, then the likelihood terms reduce to products, e.g., $p_\eta(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^N p_{\eta_i}(y_i|f(t_i))$.

Modelling Outlying Observations: A specific instance where we may wish to consider non-normal noise occurs if we believe that some of the observations $\mathbf{y}$ may be outliers. In such situations, we are required to select both the probability w of an observation y_i being an outlier and, when it is an outlier, its specific distribution. We will denote the observational noise in an outlier model as ζ. Our likelihood for the observed y_i given $f(t_i)$ then becomes a mixture:

$$p_\zeta(y_i|f_i) = (1-w)p_\epsilon(y_i|f_i) + w p_{\mathrm{o}}(y_i|f_i)$$

Here, $p_\epsilon(y_i|f_i)$ is the quoted *non-outlier* likelihood; and $p_\text{o}(x)$ the selected outlier likelihood. A natural choice for $p_\text{o}(y_i|f_i)$ may be a uniform distribution $U[f_i - a, f_i + b]$ where a and b are chosen suitably (or even simply $U[a, b]$). We can then proceed as above using the mixture $p_\zeta(x)$ as our observational model.

3.3. Additional Constraints

We are also able to incorporate additional types of constraints on the value of the function beyond simply direct observations of the function at individual times. Examples might include additional prior information, or observed information that might depend upon the value of the function at multiple times. For example, suppose that we have a belief that the gradient of the pH function should not change by more than x per million years. This can be encapsulated by modifying $\pi(\mathbf{f})$, the standard GP prior, to instead be $\pi'(\mathbf{f}) \propto \mathbb{1}_{[\max \text{gradient} < x]}\pi(\mathbf{f})$. To include this additional belief, we can simply sample from the standard (non-gradient-constrained) GP posterior, and then reject those realisations for which the maximum gradient is greater than x per million years. We note that, in practice, we estimate the maximum gradient of the function by sampling the GP extremely densely in time.

Aside: We can build up our posterior by using the GP posterior from a subset of the observations; and then use rejection sampling to adjust/update this for the full set of observations. Suppose that we observe $y_i = f(t_i) + \epsilon_i$ for $i = 1, \ldots, M$. We can sample from a reduced posterior considering all the observations excluding one, without loss of generality we suppose this is y_M:

$$g(\mathbf{f}|y_1, \ldots, y_{M-1}) \propto \prod_{i=1}^{M-1} p_\epsilon(y_i|f_i)\pi(\mathbf{f}).$$

To update this preliminary distribution to the full target posterior using all the observations, we sample from the reduced posterior $g(\mathbf{f}|y_1, \ldots, y_{M-1})$ ensuring we include the value at $f_M = f(t_M)$. We then accept this draw with probability $\frac{p_\epsilon(y_M|f_M)}{\max_f p_\epsilon(y_M|f_M)}$. Otherwise we sample from the reduced envelope again. This reduced-to-full approach is however likely to be much less efficient than sampling from an appropriate envelope based upon all the samples.

4. Rejection Sampling for $\delta^{11}\mathrm{B_{sw}}$ with Diverse Constraints

When reconstructing $\delta^{11}\mathrm{B_{sw}}$ we have multiple types of constraints: non-Gaussian observations, upper/lower bounds, and restrictions on the maximum rate of change over time. We are required to integrate all these varied constraints into our GP posterior. We illustrate how this is achieved in Figure S4. The large upper panel shows a hypothetical signal (in the thick black line) which is assumed unknown. We wish to reconstruct this function using a Gaussian Process and five noisy observations. Three observations, shown in blue, are subject to Gaussian noise (displayed at 1 standard deviation uncertainty). The other two observations, shown in green, are subject to non Gaussian noise reminiscent of a Tukey window - derived from a uniform distribution with Gaussian noise in the end members.

Here a Gaussian Process with prescribed hyperparameters (length scale of 15, noise scale of 30) assimilates the three observations with Gaussian uncertainty. This will be our envelope density. Three proposed samples are drawn from the Gaussian Process, shown in grey and labelled: **a** (dashed line), **b** (dotted line), and **c** (solid line). Looking at each sample in the upper panel, we see that:

• Gaussian Process sample **a** is completely inconsistent with the non-Gaussian observation around $t = 70$.

• Gaussian Process sample **b** is potentially consistent with both non-Gaussian observations, but is right at the limits of the possible outcomes for the constraint around $t = 30$.

• Gaussian Process sample **c** is consistent with both non-Gaussian observations - passing through a high probability region of both.

Each of the five observations is shown in a separate subpanel beneath the main time series. We might consider these as time slices through the main panel, displaying each observation probabilistically. The true value of the signal at these times is shown by the black horizontal line in each panel. The value of each Gaussian Process sample at the time slices is shown in the panels with the corresponding line style.

To assimilate observations with non-Gaussian uncertainties (shown in green), we use a rejection sampling strategy described above. To calculate a probability of acceptance, first each non-Gaussian distribution is scaled such the the maximum is equal to one (as shown in the lower panels). Then for each sample drawn from the Gaussian Process, the relative likelihood of the sample is calculated (this is shown numerically for each sample in the lower panels). The relative likelihood of each sample from the Gaussian Process is the product of the likelihoods at each of these individual timeslices.

• Gaussian Process sample **a** has a $1.0 \times 0.0 = 0$ probability of acceptance.

• Gaussian Process sample **b** has a $0.56 \times 1.0 = 0.56$ probability of acceptance.

• Gaussian Process sample **c** has a $1.0 \times 1.0 = 1.0$ probability of acceptance.

This strategy allows us to draw samples which are consistent with different types of observation - though we note there are potential failure conditions. If one of the non-Gaussian observations were much higher than the Gaussian observations, every sample would receive a 0 probability of acceptance. We mitigate against this failure condition by giving non-Gaussian constraints the possibility of being an outlier (as described above in Section 3.2).

In addition to the types of constraints shown above, we also place limitations on the rate of change in $\delta^{11}B_{sw}$. The same technique as shown for the non-Gaussian constraints is used to enforce these constraints. Using the same synthetic example, this would appear as in Figure S5.

Here the gradient in each of the samples is calculated using the first difference, and a weight for each sample can be determined by comparing this gradient to the constraints. In both the synthetic data example and the $\delta^{11}B_{sw}$ reconstruction we place a uniform prior on the gradient, which effectively describes the maximum rate of change (either in a negative or positive direction). This is displayed using horizontal bars in the large panel, within which the signal must fall, and each uniform window is plotted in individual subpanels underneath. We see that two samples remain within the imposed constraints, whereas the sample **a** is incompatible with both the earliest and latest constraint.

The prescribed maximum rate of change in both the synthetic example and the reconstruction of $\delta^{11}B_{sw}$ depends on time. In the synthetic example, the gradient is constrained in three places, with increasing acceptable range from ± 0.2 units in the earliest constraint to ± 0.6 units in the latest. For $\delta^{11}B_{sw}$, the maximum rate of change is constrained for each

discretised age window (at a resolution of 0.1Myr), and grows linearly from 0.1‰/Myr in the modern day to 1‰/Myr at 100Ma to account for increasing uncertainty in this limit.

If we run the algorithm described above for 10,000 Gaussian Process samples, accepting and rejecting the proposed samples according to the rejection algorithm, we can obtain a set of realisations from the complete posterior that incorporate all the various forms of information we have on its value: the three Gaussian observations, the two non-Gaussian, and the gradient constraints. We can then summarise these using a median and 95% pointwise posterior probability window and compare agreement to the underlying original signal as shown in Figure S6. We see a good match, within the limitations imposed by not having many observations on which to base our reconstruction, and considering that each of these observations has substantial uncertainty.

5. Data File Description

5.1. Data Supplement S1

Data Supplement S1 is an excel file containing two worksheets. The first has every accepted reconstructed $\delta^{11}B_{sw}$ time series, with age in each column, and an independent statistical sample in each row. The second worksheet contains summary metrics, specifically the median and 5% and 95% quantiles of the time series. These quantiles give a sense of uncertainty at any individual time, and can be used to propagate uncertainties when targeting absolute pH reconstructions from $\delta^{11}B_4$ within a narrow time window. When looking at longer term trends, or robustly assessing uncertainty in *change* in pH, the full time series should be integrated by sampling from the time series presented in the former tab.

5.2. Data Supplement S2

Data Supplement S3 is an excel file containing three triples of worksheets (nine in total) which contain the data, summary metrics for fits, and 10,000 Gaussian Process samples for the evolution of $^{87/86}Sr$, δ^7Li, and $^{187/188}Os$ (as shown in Figure 4 and Figure S1).

Strontium and lithium signals are taken from Misra and Froelich (2012). Osmium deserves special mention here as no Cenozoic compilation was found in an accessible format. Our Cenozoic $^{187/188}Os$ record was constructed from previously published data in Josso et al. (2019); Klemm, Levasseur, Frank, Hein, and Halliday (2005); Oxburgh (1998); Oxburgh, Pierson-Wickmann, Reisberg, and Hemming (2007); Paquay, Ravizza, Dalai, and Peucker-Ehrenbrink (2008); Paquay, Ravizza, and Coccioni (2014); Pegram and Turekian (1999); Peucker-Ehrenbrink and Ravizza (2000, 2020); van der Ploeg et al.

(2018); Ravizza (1993); Ravizza and Turekian (1992); Ravizza and Peucker-Ehrenbrink (2003); Ravizza, Norris, Blusztajn, and Aubry (2001); Reusch, Ravizza, Maasch, and Wright (1998); Robinson, Ravizza, Coccioni, Peucker-Ehrenbrink, and Norris (2009). Ages of the data from Paquay et al. (2008) were adjusted to match the age model of the record of Paquay et al. (2014). Our $^{187/188}$Os compilation integrates data from pelagic sediments and Fe-Mn crusts into a single record. The trends are broadly consistent with those previously reported in Peucker-Ehrenbrink and Ravizza (2020), however we are able to produce a representative curve with propagated uncertainties using a Gaussian Process. For most signals in this work we have used the residence time of the element in question to determine the length scale of the Gaussian process, however in the case of osmium the residence time is too short (Oxburgh, 2001) for this to be viable given the current data density. Instead we choose a low (1 Myr) but still inflated value which bridges the gaps between data without overly smoothing the signal, in order to produce the curve we believe to be most representative.

5.3. Data Supplement S3

Data Supplement S3 is an excel file containing four worksheets describing summary metrics and 10,000 possible evolutions of $\delta^{11}B_4$ and pH (as shown in Figure 3 and Figure S1).

References

Josso, P., Parkinson, I., Horstwood, M., Lusty, P., Chenery, S., & Murton, B. (2019, May). Improving confidence in ferromanganese crust age models: A composite geochemical approach. *Chemical Geology*, *513*, 108–119. doi: 10.1016/j.chemgeo.2019

.03.003

Klemm, V., Levasseur, S., Frank, M., Hein, J. R., & Halliday, A. N. (2005, September). Osmium isotope stratigraphy of a marine ferromanganese crust. *Earth and Planetary Science Letters*, *238*(1), 42–48. doi: 10.1016/j.epsl.2005.07.016

Misra, S., & Froelich, P. N. (2012, February). Lithium Isotope History of Cenozoic Seawater: Changes in Silicate Weathering and Reverse Weathering. *Science*, *335*(6070), 818–823. doi: 10.1126/science.1214697

Oxburgh, R. (1998, June). Variations in the osmium isotope composition of sea water over the past 200,000 years. *Earth and Planetary Science Letters*, *159*(3), 183–191. doi: 10.1016/S0012-821X(98)00057-0

Oxburgh, R. (2001). Residence time of osmium in the oceans. *Geochemistry, Geophysics, Geosystems*, *2*(6). doi: 10.1029/2000GC000104

Oxburgh, R., Pierson-Wickmann, A.-C., Reisberg, L., & Hemming, S. (2007, November). Climate-correlated variations in seawater 187Os/188Os over the past 200,000 yr: Evidence from the Cariaco Basin, Venezuela. *Earth and Planetary Science Letters*, *263*(3), 246–258. doi: 10.1016/j.epsl.2007.08.033

Paquay, F. S., Ravizza, G., & Coccioni, R. (2014, November). The influence of extraterrestrial material on the late Eocene marine Os isotope record. *Geochimica et Cosmochimica Acta*, *144*, 238–257. doi: 10.1016/j.gca.2014.08.024

Paquay, F. S., Ravizza, G. E., Dalai, T. K., & Peucker-Ehrenbrink, B. (2008, April). Determining Chondritic Impactor Size from the Marine Osmium Isotope Record. *Science*, *320*(5873), 214–218. doi: 10.1126/science.1152860

Pegram, W. J., & Turekian, K. K. (1999, December). The osmium isotopic composition change of Cenozoic sea water as inferred from a deep-sea core corrected for meteoritic contributions. *Geochimica et Cosmochimica Acta*, *63*(23), 4053–4058. doi: 10.1016/S0016-7037(99)00308-7

Peucker-Ehrenbrink, B., & Ravizza, G. (2000, June). The effects of sampling artifacts on cosmic dust flux estimates: A reevaluation of nonvolatile tracers (Os, Ir). *Geochimica et Cosmochimica Acta*, *64*(11), 1965–1970. doi: 10.1016/S0016-7037(99)00429-9

Peucker-Ehrenbrink, B., & Ravizza, G. E. (2020, January). Chapter 8 - Osmium Isotope Stratigraphy. In F. M. Gradstein, J. G. Ogg, M. D. Schmitz, & G. M. Ogg (Eds.), *Geologic Time Scale 2020* (pp. 239–257). Elsevier. doi: 10.1016/B978-0-12-824360-2.00008-5

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Cambridge, Mass: MIT Press.

Ravizza, G. (1993, July). Variations of the187Os/186Os ratio of seawater over the past 28 million years as inferred from metalliferous carbonates. *Earth and Planetary Science Letters*, *118*(1), 335–348. doi: 10.1016/0012-821X(93)90177-B

Ravizza, G., Norris, R. N., Blusztajn, J., & Aubry, M. P. (2001). An osmium isotope excursion associated with the Late Paleocene thermal maximum: Evidence of intensified chemical weathering. *Paleoceanography*, *16*(2), 155–163. doi: 10.1029/2000PA000541

Ravizza, G., & Peucker-Ehrenbrink, B. (2003, May). The marine 187Os/188Os record of the Eocene–Oligocene transition: The interplay of weathering and glaciation. *Earth*

and Planetary Science Letters, *210*(1), 151–165. doi: 10.1016/S0012-821X(03)00137-7

Ravizza, G., & Turekian, K. K. (1992, May). The osmium isotopic composition of organic-rich marine sediments. *Earth and Planetary Science Letters*, *110*(1), 1–6. doi: 10.1016/0012-821X(92)90034-S

Reusch, D. N., Ravizza, G., Maasch, K. A., & Wright, J. D. (1998, July). Miocene seawater 187Os/188Os ratios inferred from metalliferous carbonates. *Earth and Planetary Science Letters*, *160*(1), 163–178. doi: 10.1016/S0012-821X(98)00082-X

Robinson, N., Ravizza, G., Coccioni, R., Peucker-Ehrenbrink, B., & Norris, R. (2009, May). A high-resolution marine 187Os/188Os record for the late Maastrichtian: Distinguishing the chemical fingerprints of Deccan volcanism and the KP impact event. *Earth and Planetary Science Letters*, *281*(3), 159–168. doi: 10.1016/j.epsl.2009.02.019

van der Ploeg, R., Selby, D., Cramwinckel, M. J., Li, Y., Bohaty, S. M., Middelburg, J. J., & Sluijs, A. (2018, July). Middle Eocene greenhouse warming facilitated by diminished weathering feedback. *Nature Communications*, *9*(1), 2877. doi: 10.1038/s41467-018-05104-9

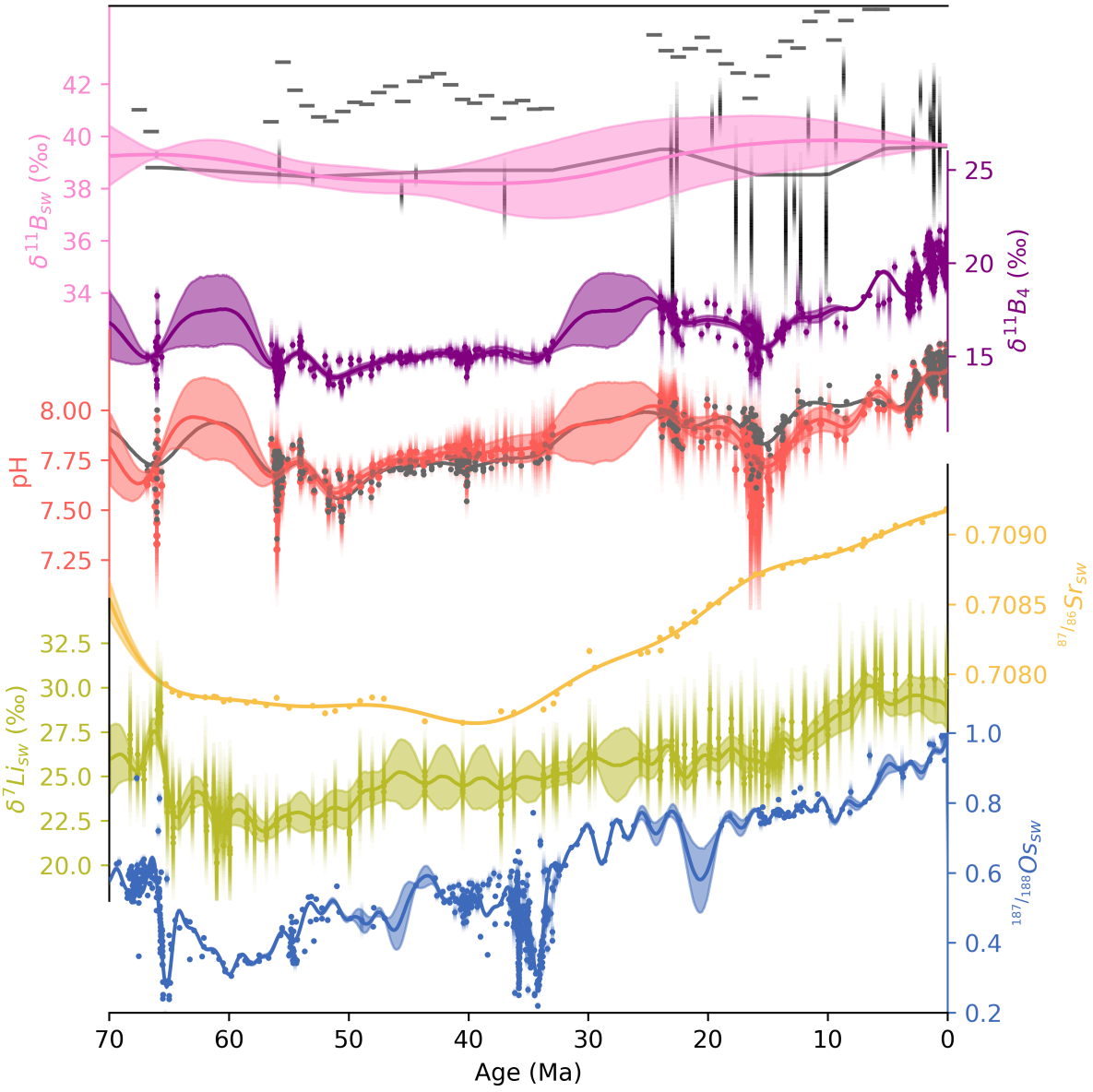

Figure S1: $\delta^{11}B_{sw}$ (pink), $\delta^{11}B_4$ (purple), pH (red), $^{87}/_{86}Sr$ (yellow), δ^7Li (green), and $^{187}/_{188}Os$ (blue) are shown here with the same style as shown in the two separate plots in the main text. We provide an large summary figure here for easy comparison of the six signals.

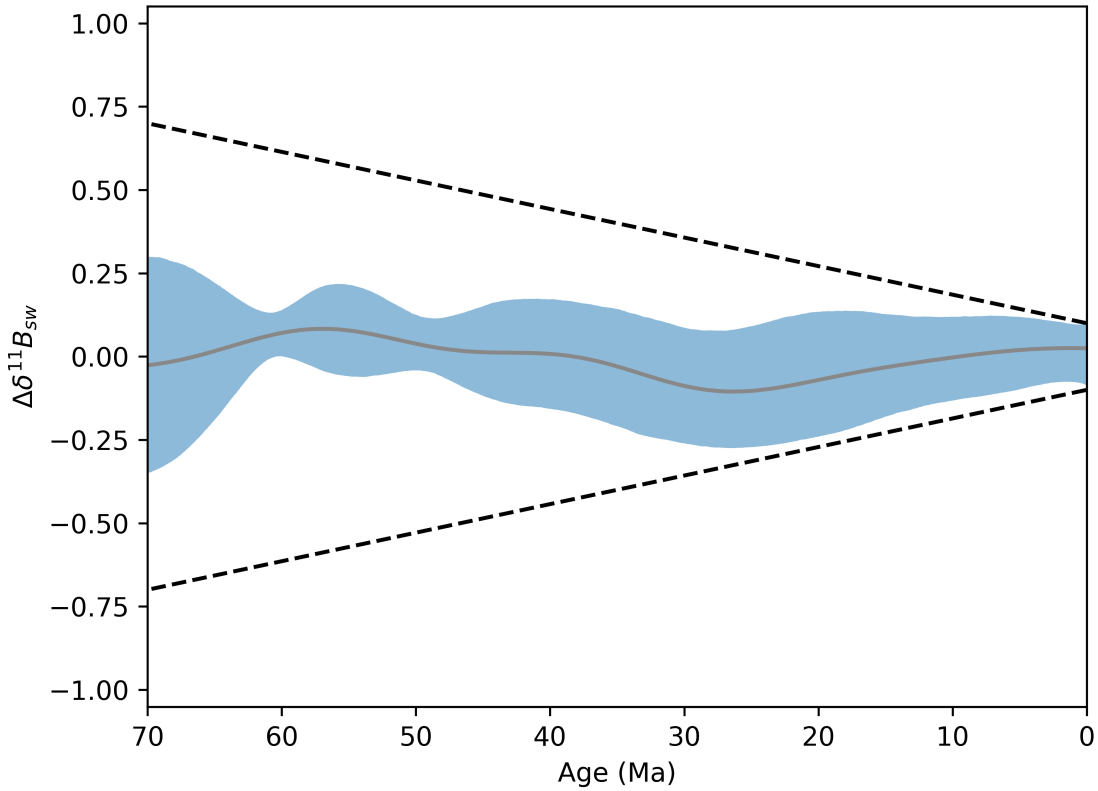

Figure S2: The temporal gradient in our $\delta^{11}B_{sw}$ samples is shown by the blue window, with the mean average shown in grey. Our imposed limitation on the gradient of $\delta^{11}B_{sw}$ through time is shown by the dotted black lines. Any sample drawn outside of these bounds would be rejected. It can be seen that the limitations have most influence between the Neogene and modern, and further back do not result in rejection of any samples.

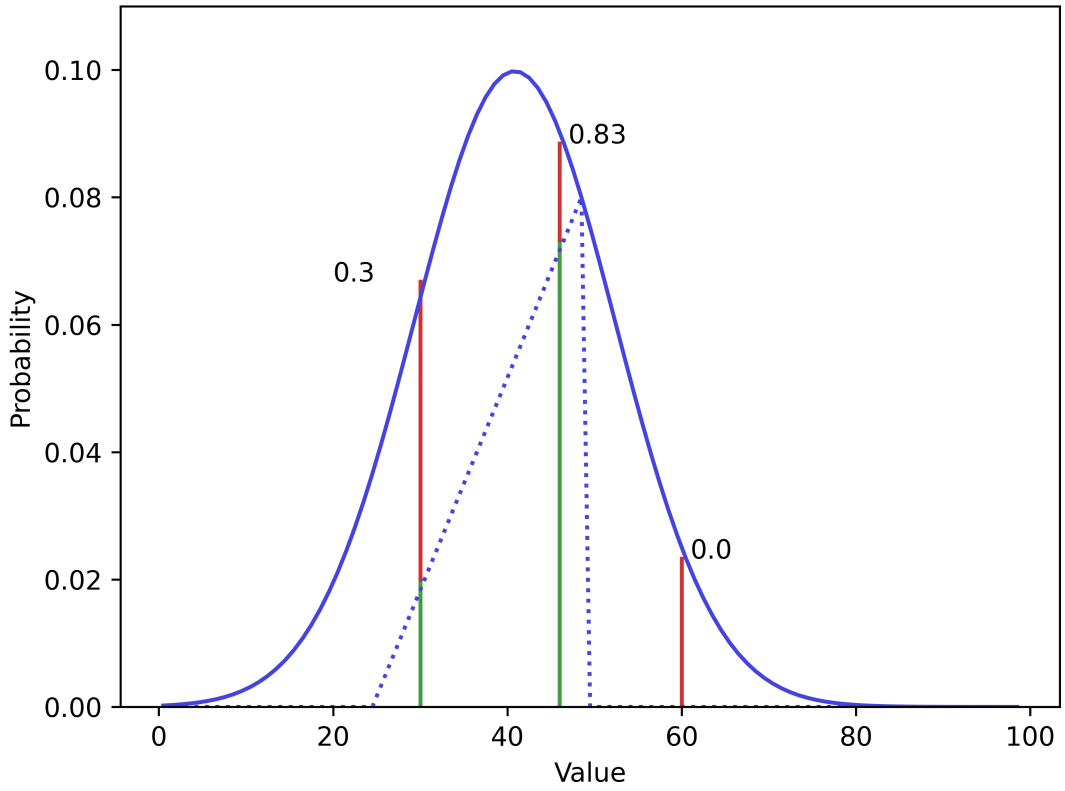

Figure S3: An illustration of rejection sampling. We aim to sample from the dotted blue saw-tooth density (shown as a dotted blue line) using a Gaussian distribution as the envelope (shown in solid blue). The Gaussian envelope has been rescaled from a standard Gaussian distribution so it encapsulates the target sawtooth density. To obtain a sample from the sawtooth distribution, we first sample a value envelope density. We show hypothetical three values $z_1 = 30$, $z_2 = 45$ and $z_3 = 60$. The probability of accepting each sample z_i as a draw from the sawtooth distribution is then the ratio of the height of the target (dotted line) compared to the height of the envelope (solid blue line). These probabilities are 0.3, 0.83, and 0 respectively for our three hypothetical z_i samples.

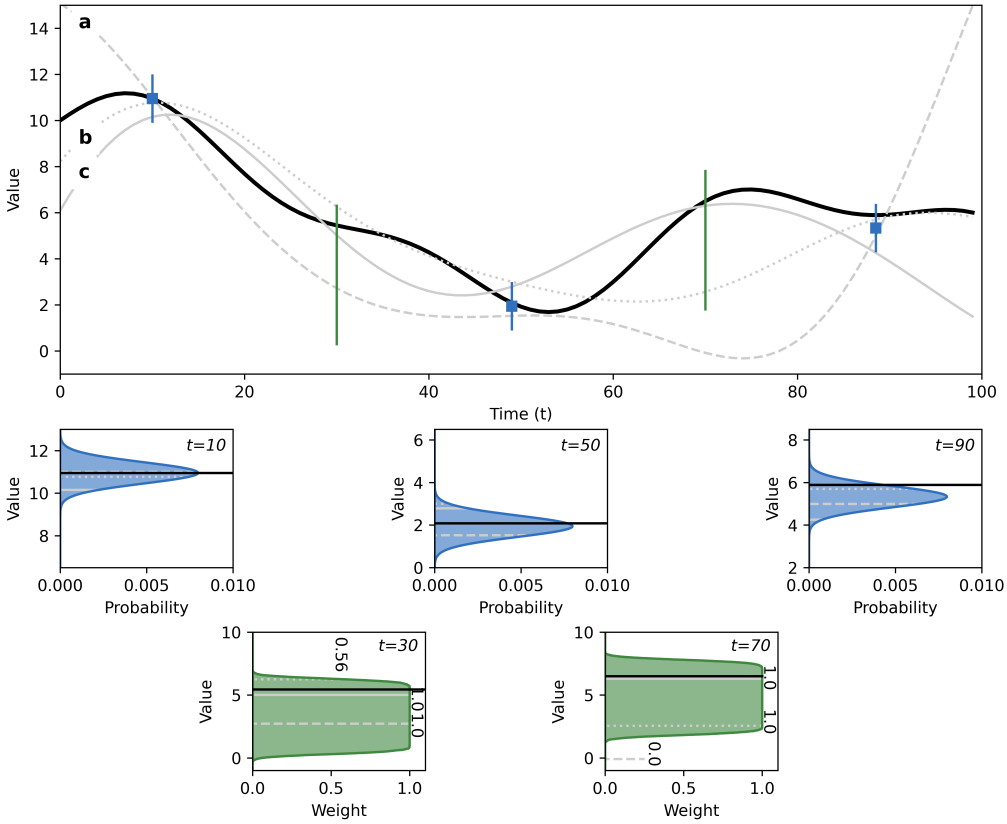

Figure S4: Calculating the GP posterior by rejection sampling that combines multiple constraints of both Gaussian and non-Gaussian types. Multiple samples are taken using a Gaussian Process conditioned on only the Gaussian constraints (samples are shown in the grey lines), and for each we quantify the probability of that sample at each data constraint (the blue and green windows). The probability of each sample is the product of the probabilities of that sample at each data constraint, meaning that sample a is rejected (it does not match the fourth data constraint), while other are likely to be accepted. This is described further in Section 3.3.

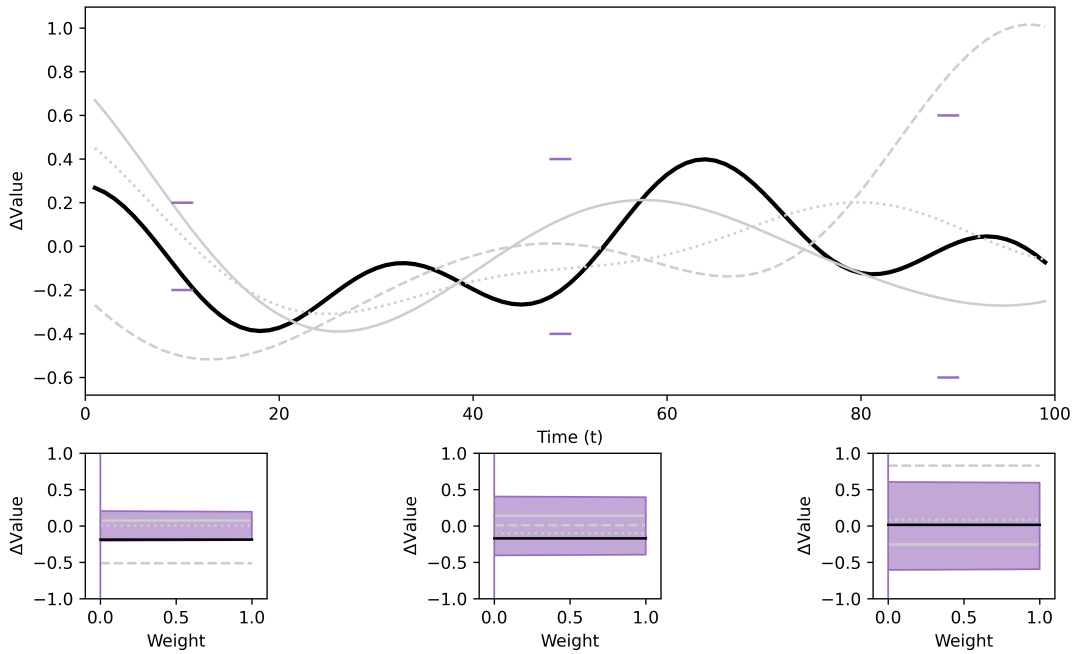

Figure S5: Incorporating constraints on the gradient into estimation of $\delta^{11}B_{sw}$. The main plot shows the estimated gradient ΔValue of our function over time. The subplots shown the gradient constraint we impose upon the signal, and the probability of observing each statistical sample at that time. Note that here we impose gradient constraint only a three discrete locations, whereas in the main text we apply a continuous limitation on the rate of change in $\delta^{11}B_{sw}$.

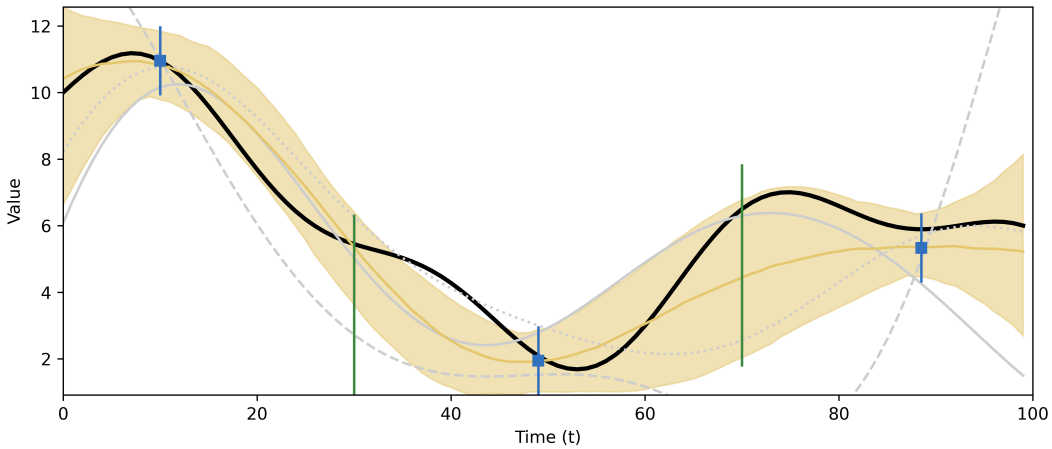

Figure S6: Reconstruction of the function shown in black incorporating information from noisy Gaussian observations (in blue) and non-Gaussian observations (in green) and gradient restrictions (in purple in Figure S5). The yellow line shows our central estimate, with a 95% confidence interval shown in the yellow shaded region.