Statistical Model Data Comparison

Fran Bragg, Phillip Paine, Alan Saul, Dan Lunt, Andrew Zammit-Mangion, **Richard Wilkinson**

School of Maths and Statistics
University of Sheffield

June 2018

Model data comparison

Palaeoclimate data can provide us with insight into the dynamics of the Earth system under conditions different from modern.

- data allow us to test our theoretical understanding of the climate through evaluation of Earth system models
- palaeoclimate data are often 'out of sample' - not used in the tuning or development of the models, and so provide a fairer test of model performance.
- Good performance in predicting past climate leads to confidence in future model prediction, and poor performance can be used to highlight deficiencies in models

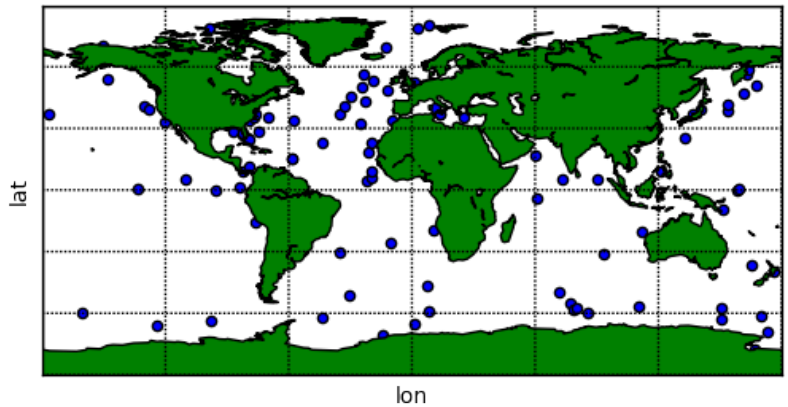
Thus, careful data-model comparison is of great importance in the verification of palaeoclimate simulations, particularly if used to rank, or even reject, models or model simulations.

Palaeodata

Palaeoclimate data is different from the modern equivalent.

- datasets of observations are often globally sparse and simultaneously clustered in certain regions,

- e.g. PRISM Pliocene sea surface temperature (SST) dataset (Dowsett et al., 2010) has a preponderance of data in the North Atlantic, and little or none in the subtropical South Pacific.

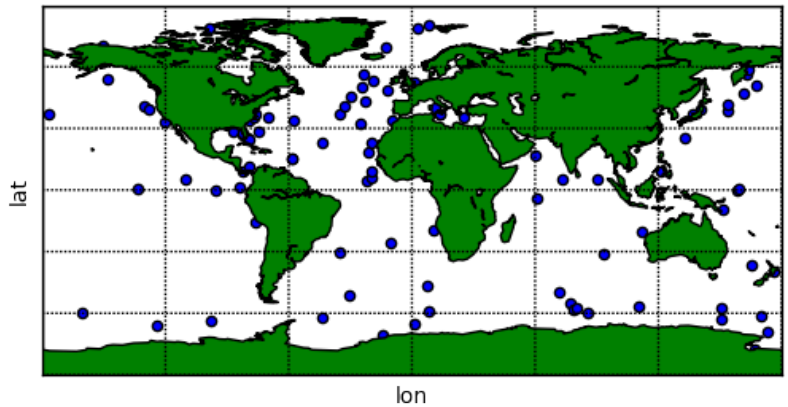


Palaeodata

Palaeoclimate data is different from the modern equivalent.

- datasets of observations are often globally sparse and simultaneously clustered in certain regions,

- e.g. PRISM Pliocene sea surface temperature (SST) dataset (Dowsett et al., 2010) has a preponderance of data in the North Atlantic, and little or none in the subtropical South Pacific.



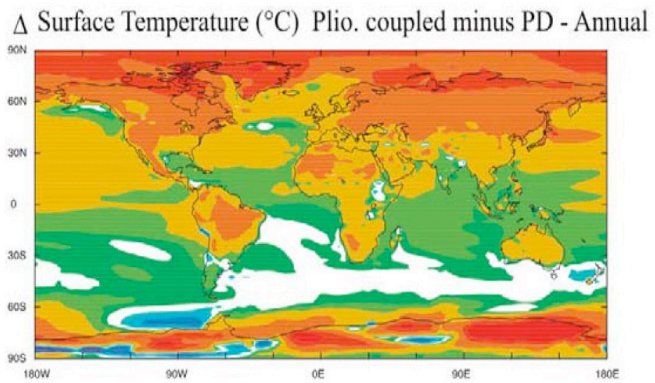
- Palaeoclimate data usually has high levels of observational uncertainty (usually poorly specified) which are often significant in relation to the climate signal being assessed.

Thus, the usual qualitative metrics developed in modern climate assessments are often inappropriate

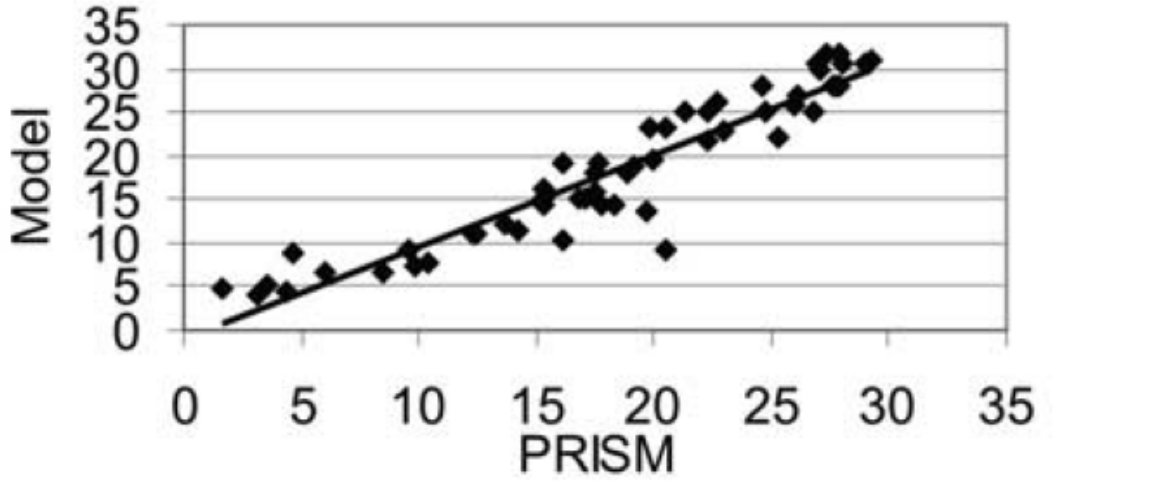
- typically rely on observations with gridded, global coverage and do not account for the uncertainties in those observations (Gleckler et al., 2008).

Current approach - Dan Lunt PEN talk 2015

Model simulation involves changing CO_2, mountain height, vegetation, and ice sheet extent.



Cross-plots of model temperatures vs. data, and RMS score.



"The results demonstrate that, as a whole, the SSTs predicted by the model and those reconstructed from geological data are not statistically different (at a 95% statistical significance level). An R^2 value of 0.88 ... suggests that the two SST data sets are particularly similar."

Haywood and Valdes, 2007

Our approach

Want a robust method that is statistically sound, yet simple to apply and understand, and which can be applied by the palaeoclimate community without a statistician.

Our approach

Want a robust method that is statistically sound, yet simple to apply and understand, and which can be applied by the palaeoclimate community without a statistician. Ideally, it should

- account for clustering,
  - a statistical representation of the data is needed.
  - use a Gaussian process approach to estimate spatial correlations
  - we can control correlations in each direction
- account for uncertainty
  - Use scoring rules to evaluate model simulations relative to the data-derived field.
  - incorporate different (and uncertain) levels of measurement uncertainty.

Our approach

Want a robust method that is statistically sound, yet simple to apply and understand, and which can be applied by the palaeoclimate community without a statistician. Ideally, it should

- account for clustering,
  - a statistical representation of the data is needed.
  - use a Gaussian process approach to estimate spatial correlations
  - we can control correlations in each direction
- account for uncertainty
  - Use scoring rules to evaluate model simulations relative to the data-derived field.
  - incorporate different (and uncertain) levels of measurement uncertainty.

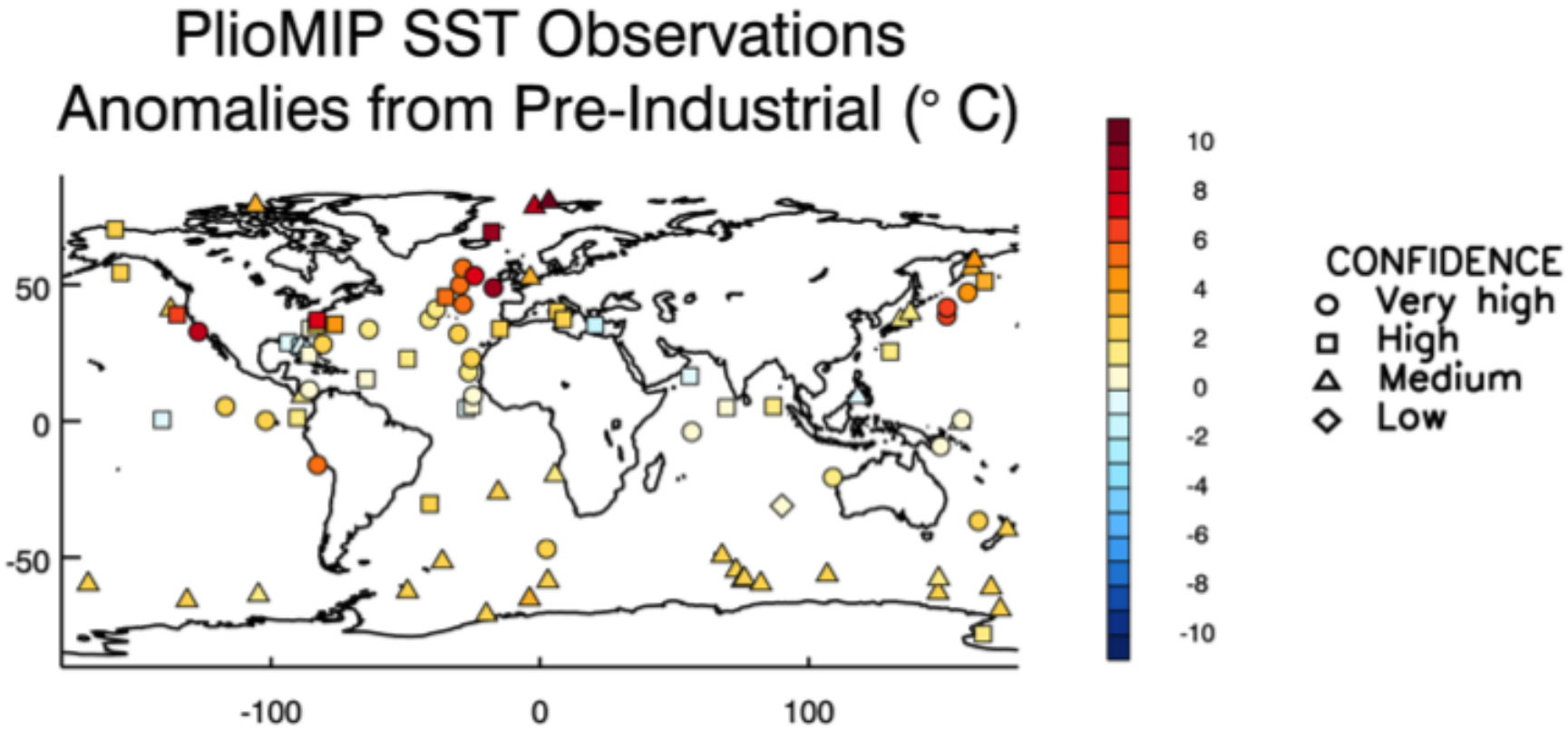
Our approach builds on Zammit-Mangion *et al.* 2014, but instead of using labour and computer intensive Gauss-Markov random fields, we use a simpler (and equivalent) Gaussian process approach.

All software available on github.

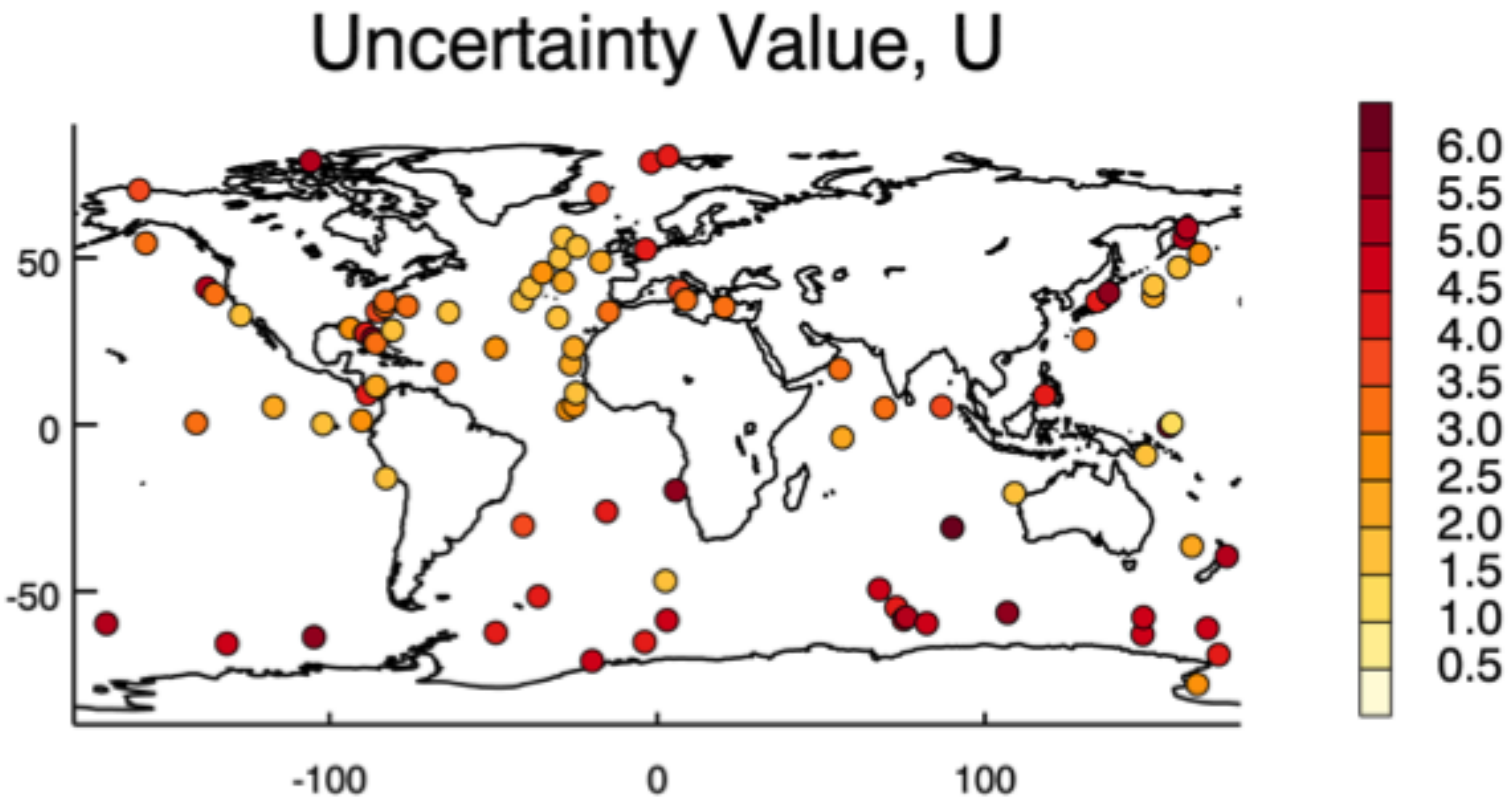
Exemplar: Scoring HadCM3 Pliocene predictions

Extended PRISM3 SST observations

Shown as anomalies from the pre-industrial HadISST dataset.

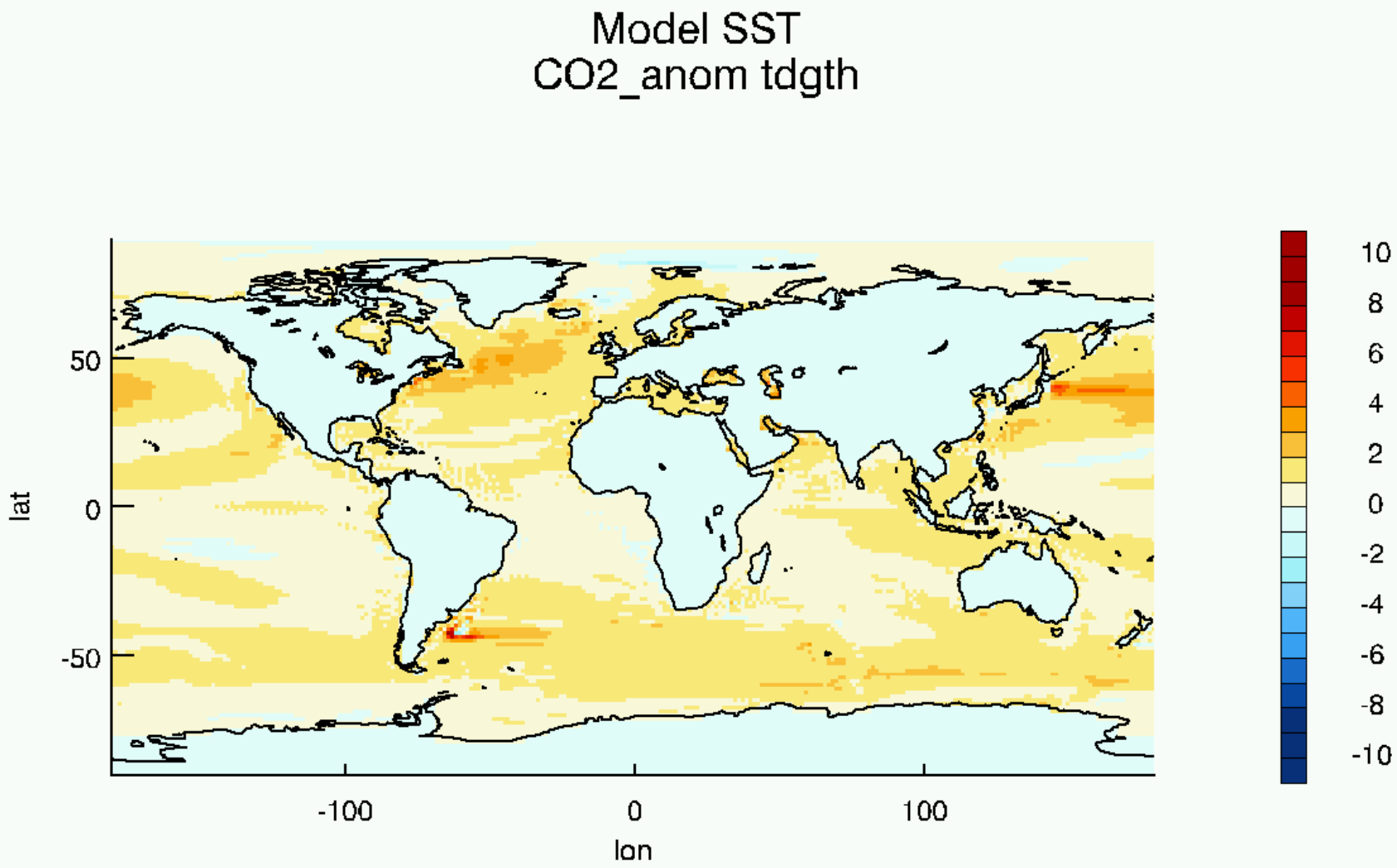


Reconstruction/measurement errors



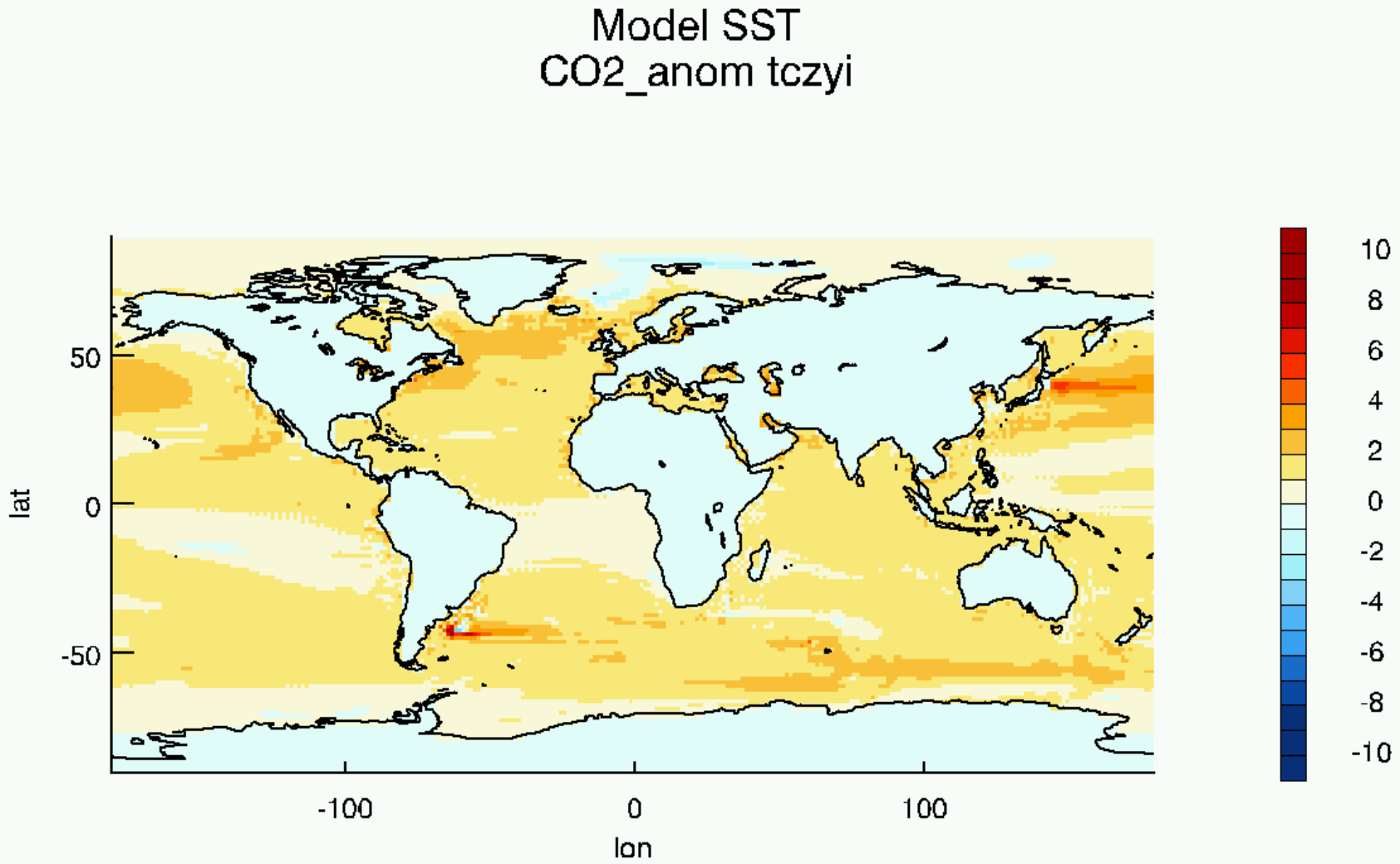
- Qualitative assessment of data quality at each site based on five criteria, giving a score between 5 and 18 (Dowsett et al., 2012).
- We converted this to a numeric estimate of the uncertainty - to be considered as a numerical construct to represent plausible uncertainty values in the absence of any estimates.
- Absolute values are not relevant, only the ratio between values.

280ppm

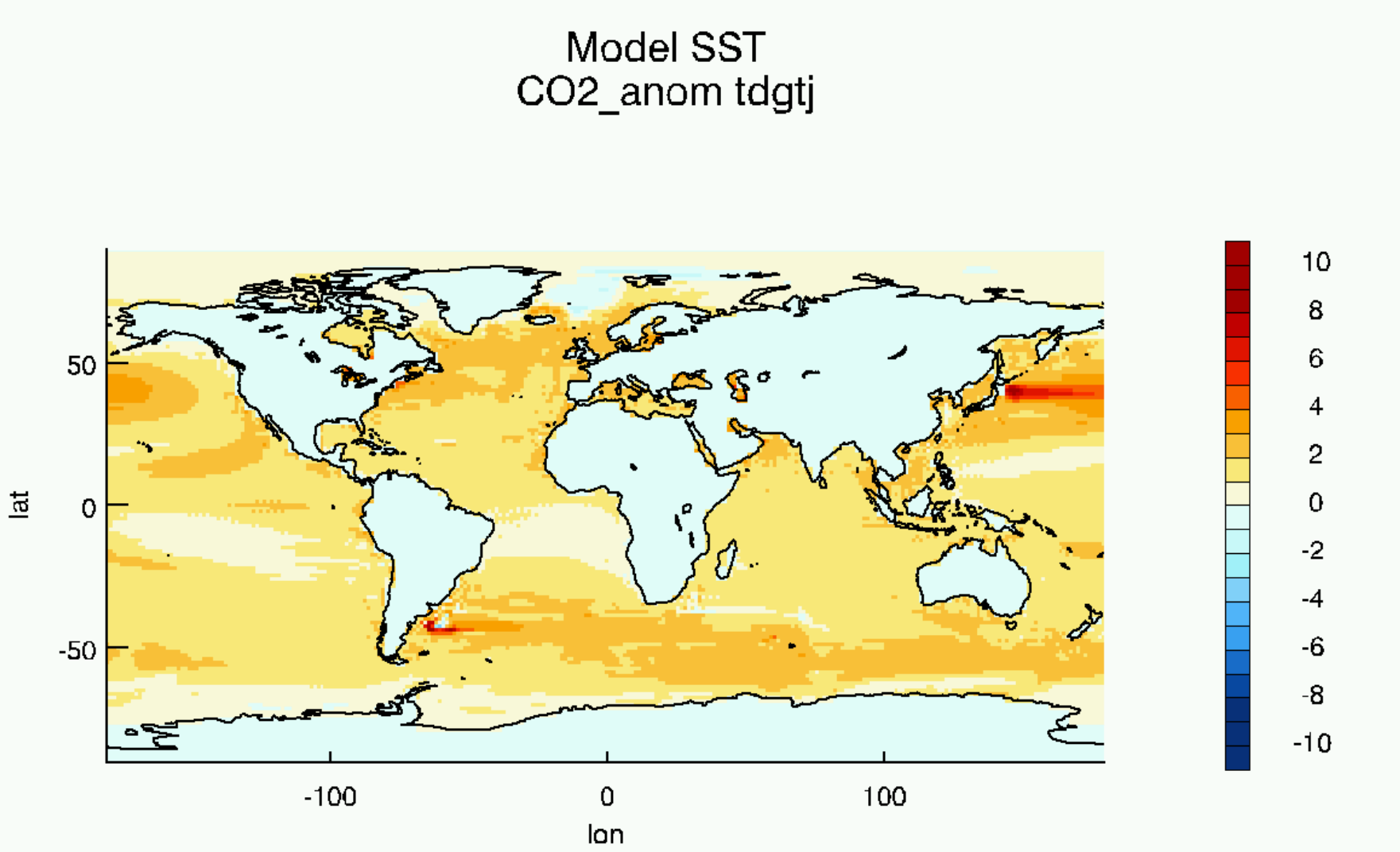


HadCM3 (Valdes et al., 2017) that follow the PlioMIP Experiment 2 protocol These simulations are identical except for being forced with a range of pCO2 levels

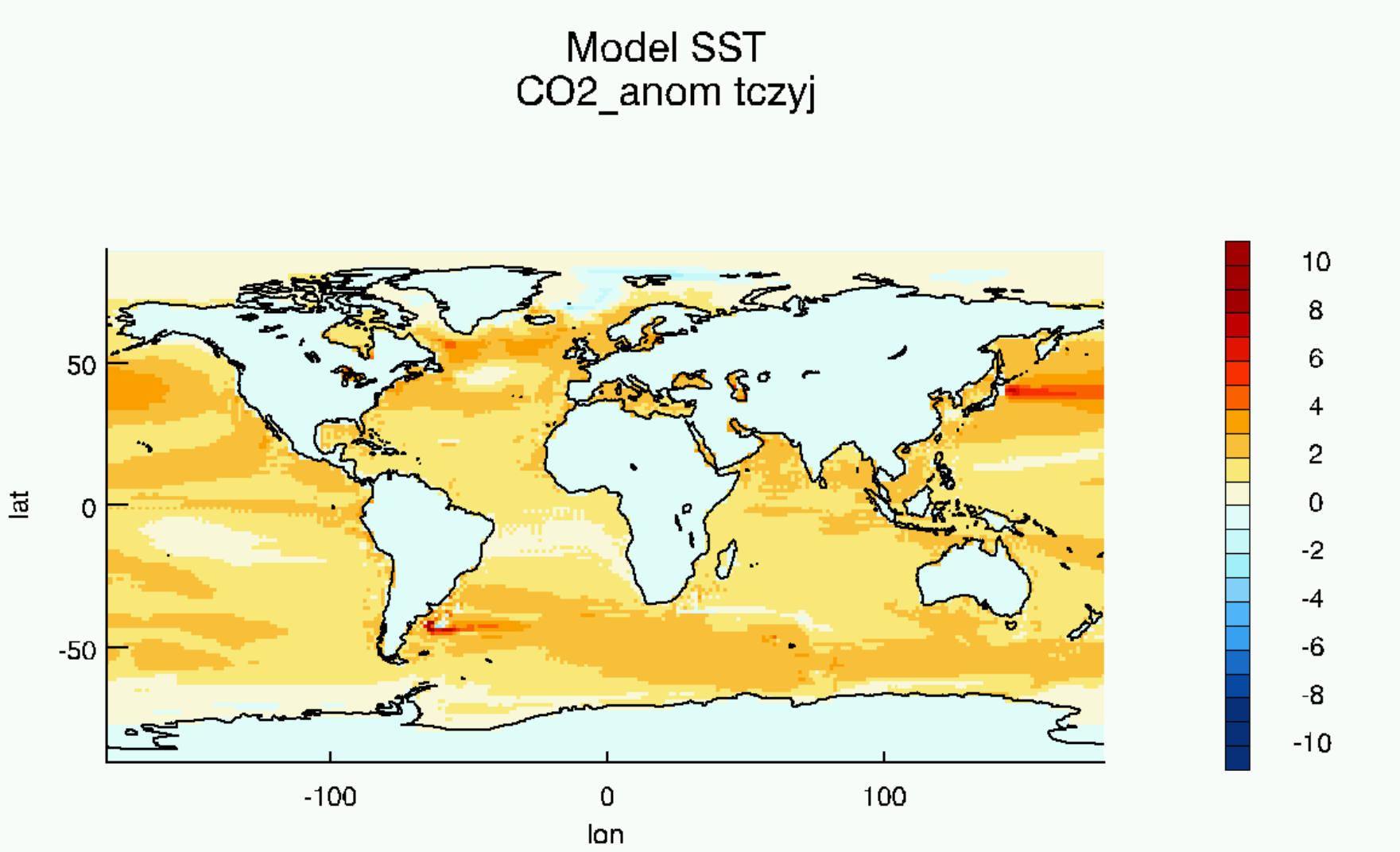
315ppm


Model SST
CO2_anom tczyi
lat
50
0
-50
-100
0
100
lon
10
8
6
4
2
0
-2
-4
-6
-8
-10

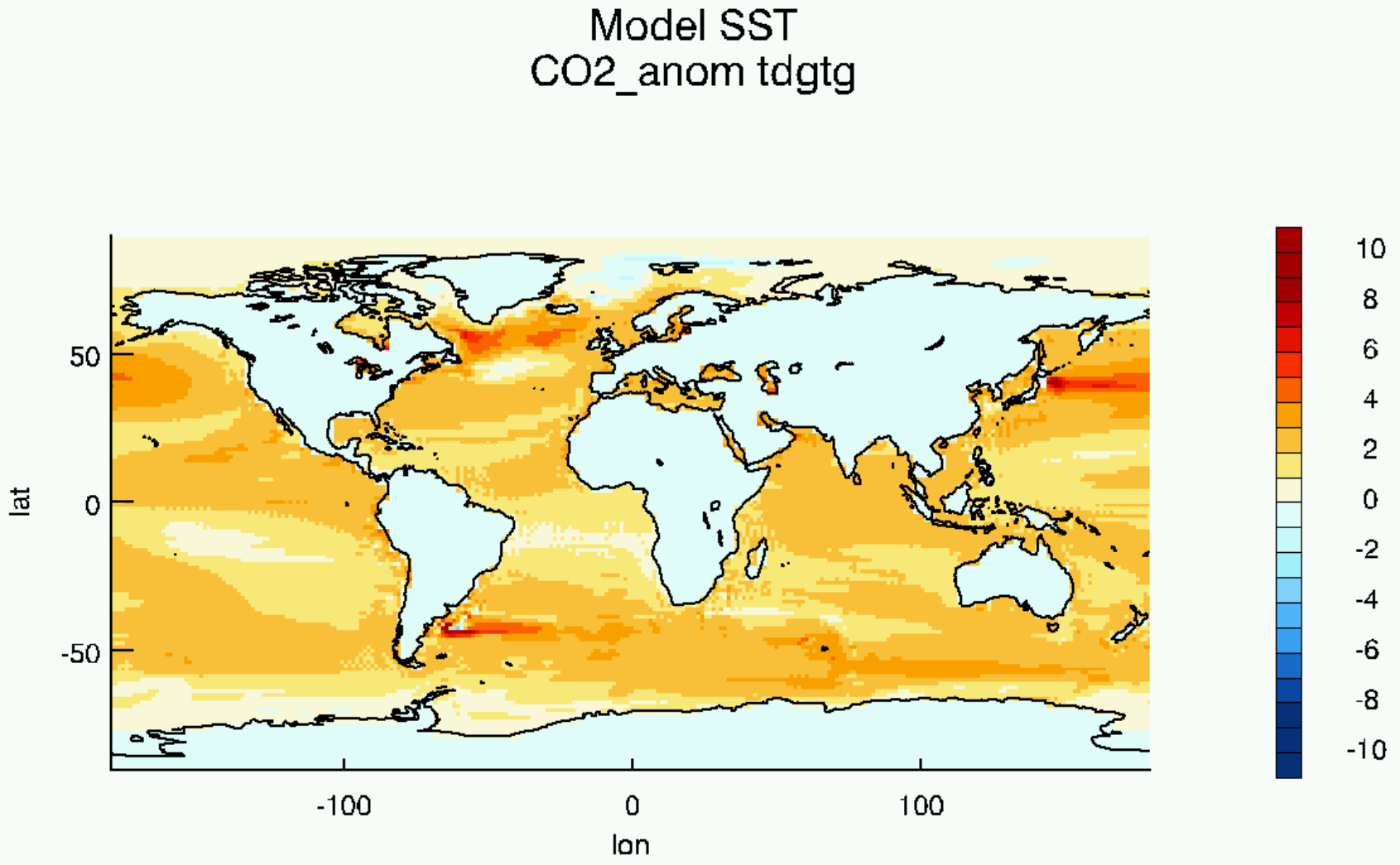
350ppm


Model SST
CO2_anom tdgtj
lat
50
0
-50
-100
0
100
lon
10
8
6
4
2
0
-2
-4
-6
-8
-10

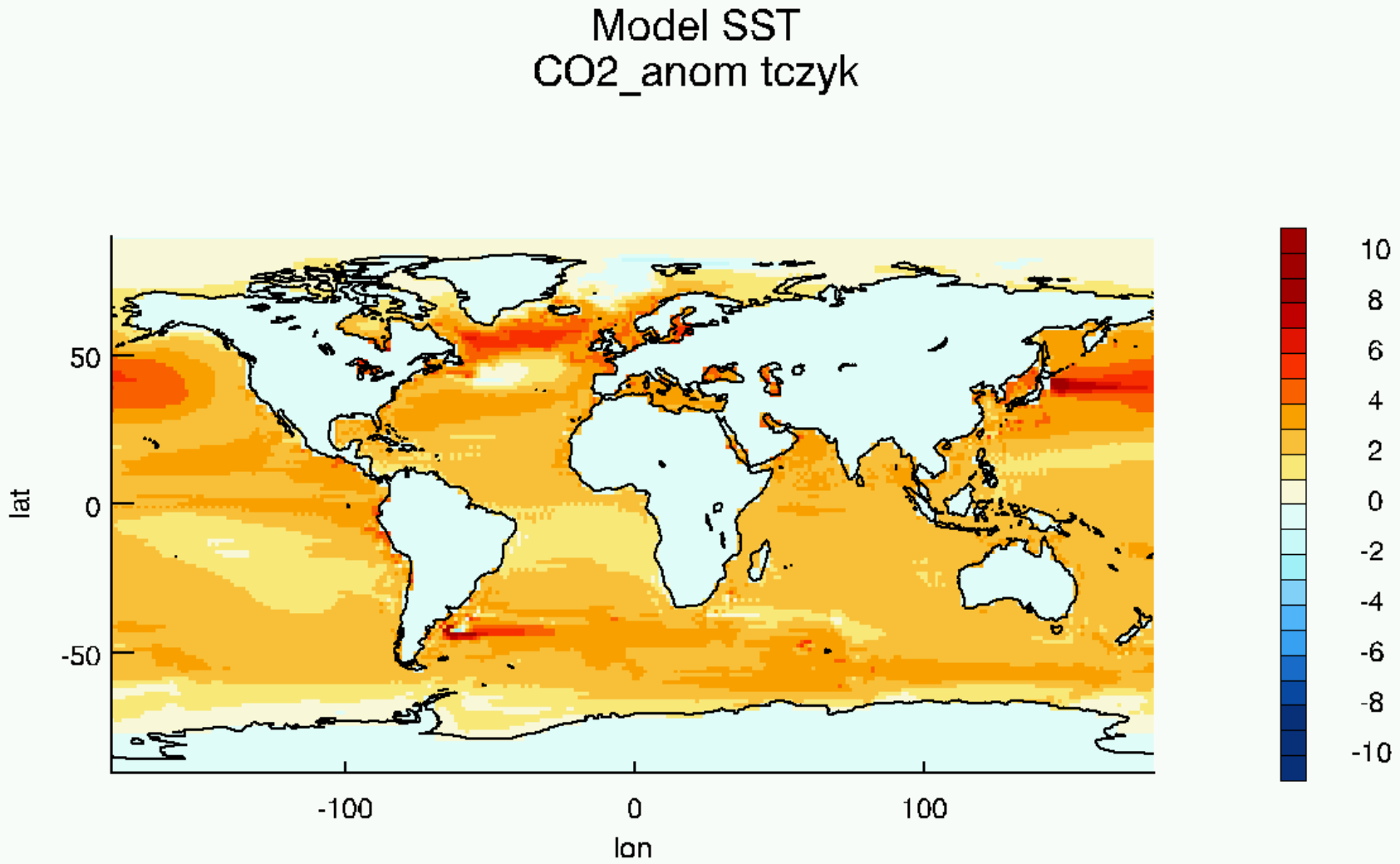
375ppm


Model SST
CO2_anom tczyj
lat
50
0
-50
-100
0
100
lon
10
8
6
4
2
0
-2
-4
-6
-8
-10

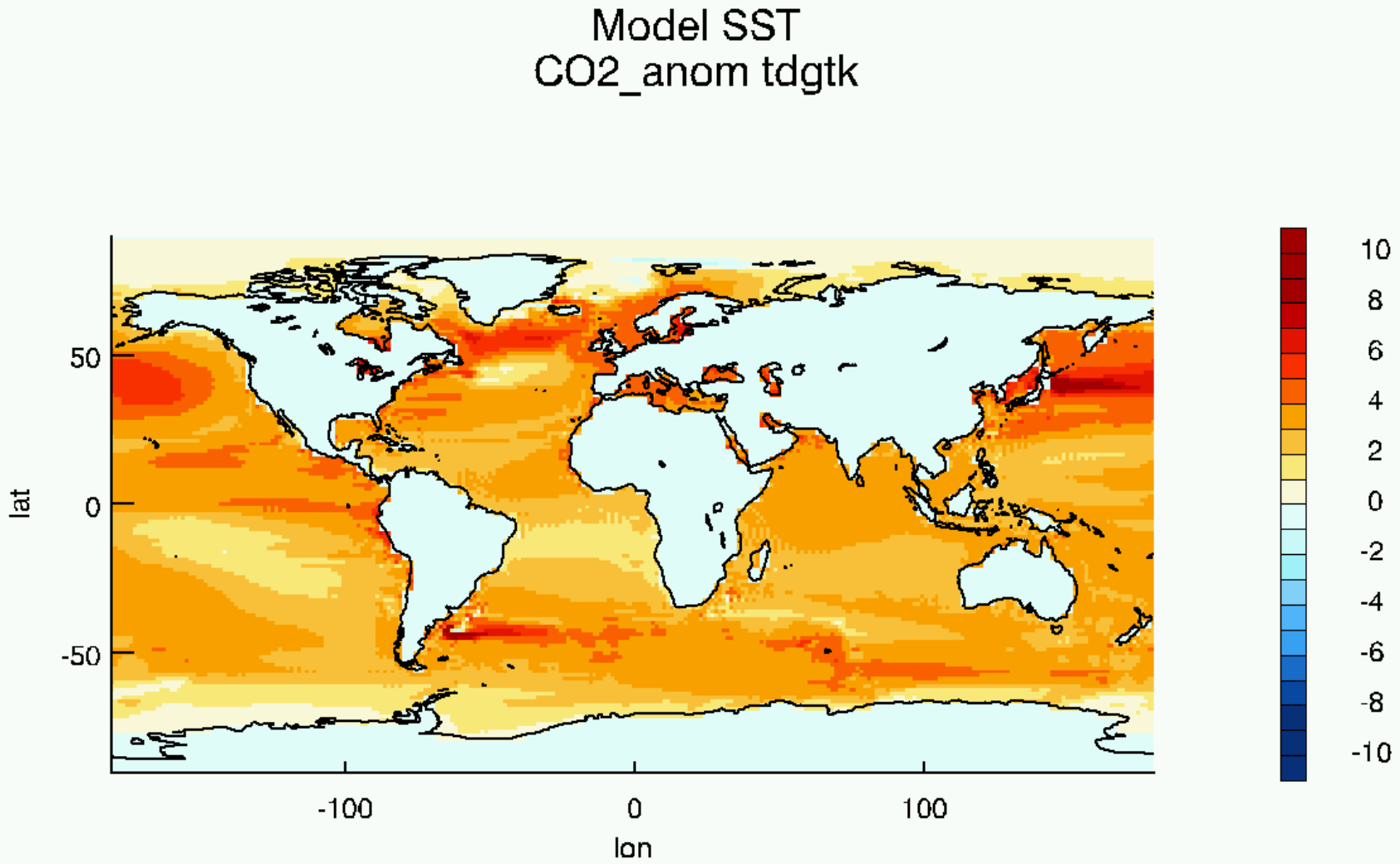
405ppm


Model SST
CO2_anom tdgtg
lat
50
0
-50
-100
0
100
lon
10
8
6
4
2
0
-2
-4
-6
-8
-10

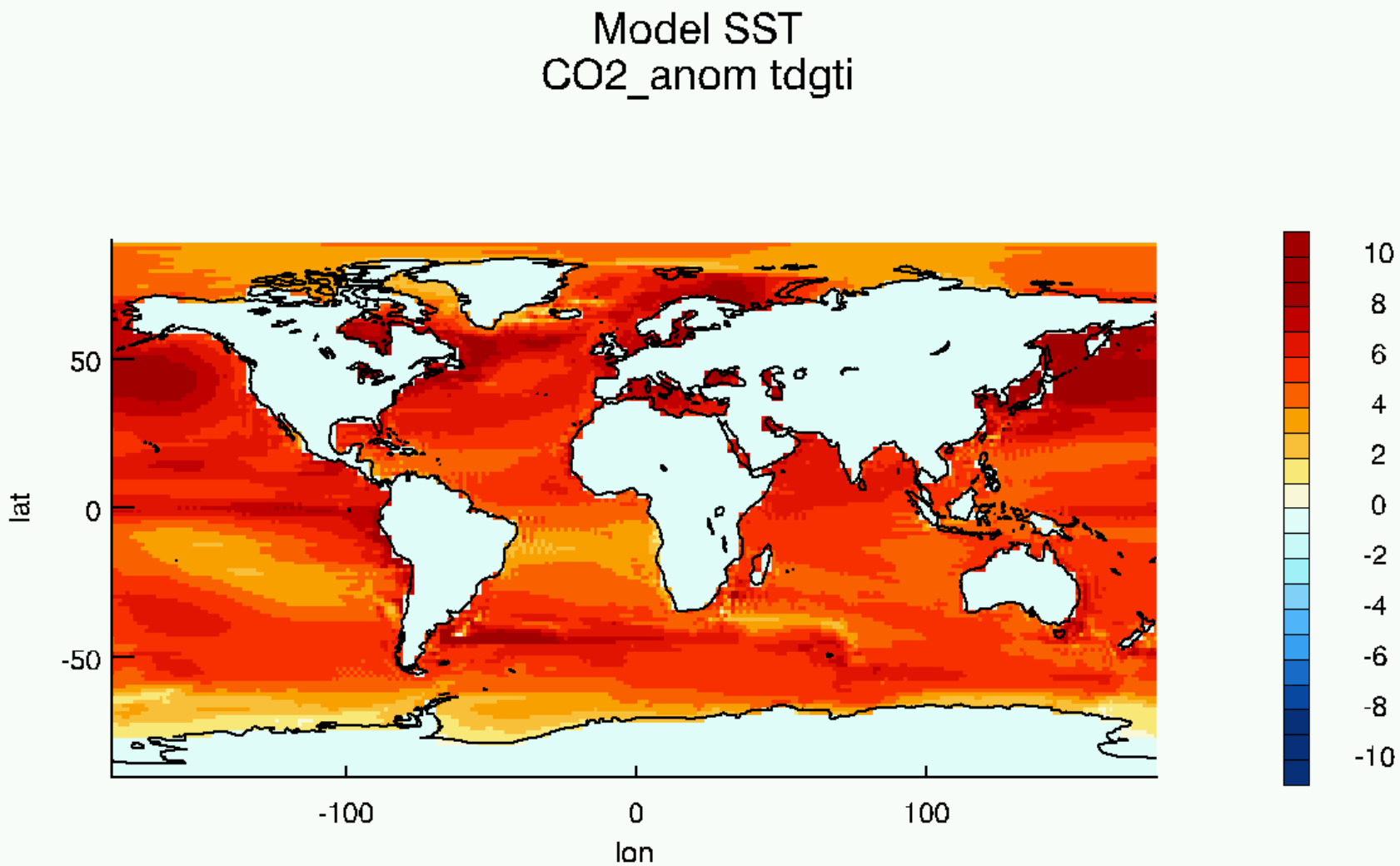
475ppm


Model SST
CO2_anom tczyk
lat
50
0
-50
-100
0
100
lon
10
8
6
4
2
0
-2
-4
-6
-8
-10

560ppm


Model SST
CO2_anom tdgtk
lat
50
0
-50
-100
0
100
lon
10
8
6
4
2
0
-2
-4
-6
-8
-10

1000ppm



Which simulation is closest to the data?

What do the data and model suggest the pCO2 level was?

Scoring rules

Suppose a forecaster makes a prediction in the form of a probability distribution P about some quantity.

We wait and observe this quantity, recording value d. How good was the prediction?

Scoring rules

Suppose a forecaster makes a prediction in the form of a probability distribution P about some quantity.

We wait and observe this quantity, recording value d. How good was the prediction?

A score maps from a distribution and observations to $\mathbb{R}$, that is $S : \mathcal{P} \times \mathbb{R}^d \to \mathbb{R}$, i.e., gives us a score

$$S(P, d)$$

Scoring rules

Suppose a forecaster makes a prediction in the form of a probability distribution P about some quantity.

We wait and observe this quantity, recording value d. How good was the prediction?

A score maps from a distribution and observations to $\mathbb{R}$, that is $S : \mathcal{P} \times \mathbb{R}^d \to \mathbb{R}$, i.e., gives us a score

$$S(P, d)$$

S is a proper score if

$$\arg\max_P \mathbb{E}_{D \sim Q}[S(P, D)] = Q$$

e.g. log likelihood

$$S(P, d) = \log p(d)$$

where p is the pdf corresponding to P.

Problem

Suppose we are given model outputs

$$M_1, \ldots, M_k$$

e.g., the SST fields from running a climate model with k different boundary conditions.

These are not probabilistic predictions. Given data D how can we score them? How do we get uncertainty into this story?

Problem

Suppose we are given model outputs

$$M_1, \ldots, M_k$$

e.g., the SST fields from running a climate model with k different boundary conditions.

These are not probabilistic predictions. Given data D how can we score them? How do we get uncertainty into this story?

Idea: Fit a statistical model to the data, $P_d(x)$, and score the model evaluations (Zammit-Mangion *et al.* 2014).

$$S(P_d, M_i)$$

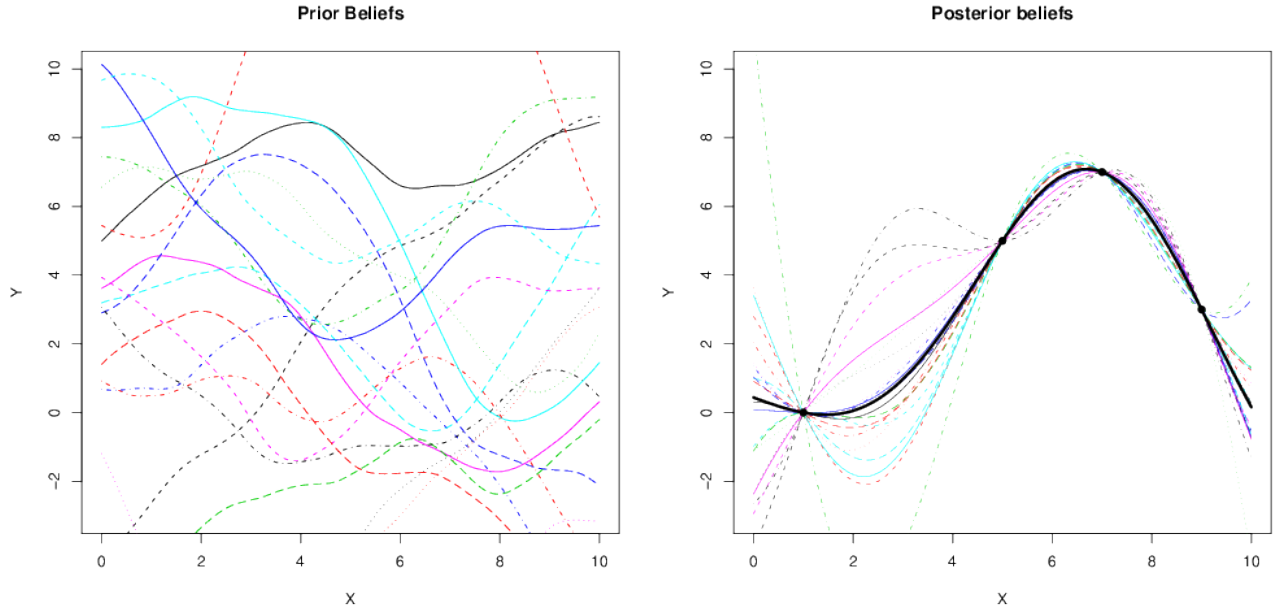
As long as P_d is sufficiently sophisticated, this will allow us to account for

- clustering in the data
- varying uncertainty levels in the data

Note the philosophically backwards application of scoring rules.

Gaussian process interpolation

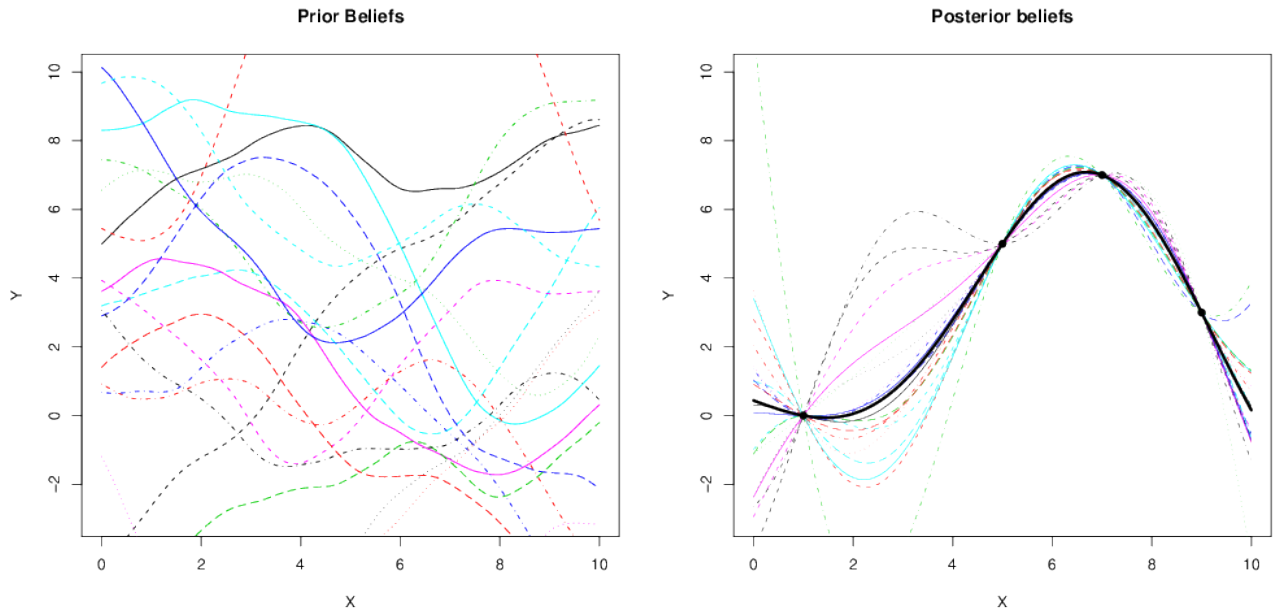
We have noisy observations D_i at locations x_i for $i = 1, \ldots, n$
Natural to use kriging to interpolate underlying surface $D(x)$



Think of these as probabilistic models of functions (or interpolation with uncertainty).

Gaussian process interpolation

We have noisy observations D_i at locations x_i for $i = 1, \ldots, n$
Natural to use kriging to interpolate underlying surface $D(x)$



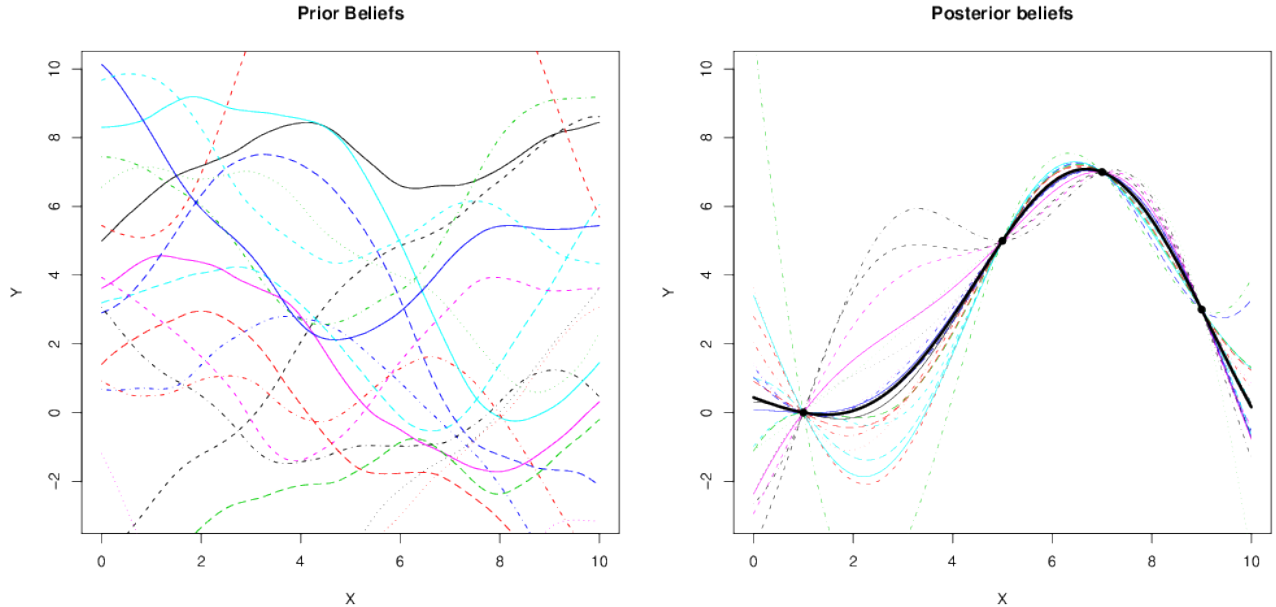
Think of these as probabilistic models of functions (or interpolation with uncertainty).
A GP is a random process indexed by $x \in \mathcal{X}$ say, such that for every finite set of indices, $x_1, \ldots, x_n$,

$$\mathbf{f} = (f(x_1), \ldots, f(x_n))$$

has a multivariate Gaussian distribution.

Gaussian process interpolation

We have noisy observations D_i at locations x_i for $i = 1, \ldots, n$
Natural to use kriging to interpolate underlying surface $D(x)$



Think of these as probabilistic models of functions (or interpolation with uncertainty).
A GP is a random process indexed by $x \in \mathcal{X}$ say, such that for every finite set of indices, $x_1, \ldots, x_n$,

$$\mathbf{f} = (f(x_1), \ldots, f(x_n))$$

has a multivariate Gaussian distribution. Why use GPs?

Answer 1

Class of models is closed under various operations.

Answer 1

Class of models is closed under various operations.

- Closed under addition

$$f_1(\cdot), f_2(\cdot) \sim GP \quad \text{then} \quad (f_1 + f_2)(\cdot) \sim GP$$

Answer 1

Class of models is closed under various operations.

- Closed under addition

$$f_1(\cdot), f_2(\cdot) \sim GP \quad \text{then} \quad (f_1 + f_2)(\cdot) \sim GP$$

- Closed under Bayesian conditioning, i.e., if we observe

$$\mathbf{D} = (f(x_1), \ldots, f(x_n))$$

then

$$f|D \sim GP$$

but with updated mean and covariance functions.

Answer 1

Class of models is closed under various operations.

- Closed under addition

$$f_1(\cdot), f_2(\cdot) \sim GP \quad \text{then} \quad (f_1 + f_2)(\cdot) \sim GP$$

- Closed under Bayesian conditioning, i.e., if we observe

$$\mathbf{D} = (f(x_1), \ldots, f(x_n))$$

then

$$f|D \sim GP$$

but with updated mean and covariance functions.

- Closed under any linear operation. If $\mathcal{L}$ is a linear operator, then

$$\mathcal{L}f \sim GP(\mathcal{L}m, \mathcal{L}k\mathcal{L}^\top)$$

e.g. $\frac{df}{dx}$, $\int f(x)dx$, Af are all GPs

Answer 2: non-parametric/kernel regression

- Linear regression $y = x^\top \beta + \epsilon$ can be written solely in terms of inner products $x^\top x$.

$$\begin{aligned}\hat{\beta} &= \arg\min ||y - X\beta||_2^2 + \sigma^2 ||\beta||_2^2 \\ &= X^\top (XX^\top + \sigma^2 I)^{-1} y \quad \text{(the dual form)}\end{aligned}$$

Answer 2: non-parametric/kernel regression

- Linear regression $y = x^\top \beta + \epsilon$ can be written solely in terms of inner products $x^\top x$.

$$\begin{aligned}\hat{\beta} &= \arg\min ||y - X\beta||_2^2 + \sigma^2 ||\beta||_2^2 \\ &= X^\top (XX^\top + \sigma^2 I)^{-1} y \quad \text{(the dual form)}\end{aligned}$$

- We know that we can replace x by a feature vector in linear regression, e.g., $\phi(x) = (1\ x\ x^2\ \cos(x))^\top$ etc.

Answer 2: non-parametric/kernel regression

- Linear regression $y = x^\top \beta + \epsilon$ can be written solely in terms of inner products $x^\top x$.

$$\begin{aligned}\hat{\beta} &= \arg\min ||y - X\beta||_2^2 + \sigma^2 ||\beta||_2^2 \\ &= X^\top (XX^\top + \sigma^2 I)^{-1} y \quad \text{(the dual form)}\end{aligned}$$

- We know that we can replace x by a feature vector in linear regression, e.g., $\phi(x) = (1\ x\ x^2\ \cos(x))^\top$ etc.
- For some features, inner product is equivalent to evaluating a kernel

$$\phi(x)^\top \phi(x') \equiv k(x, x')$$

where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a semi-positive definite function.

Answer 2: non-parametric/kernel regression

- Linear regression $y = x^\top \beta + \epsilon$ can be written solely in terms of inner products $x^\top x$.

$$\begin{aligned}\hat{\beta} &= \arg\min ||y - X\beta||_2^2 + \sigma^2 ||\beta||_2^2 \\ &= X^\top (XX^\top + \sigma^2 I)^{-1} y \quad \text{(the dual form)}\end{aligned}$$

- We know that we can replace x by a feature vector in linear regression, e.g., $\phi(x) = (1\ x\ x^2\ \cos(x))^\top$ etc.
- For some features, inner product is equivalent to evaluating a kernel

$$\phi(x)^\top \phi(x') \equiv k(x, x')$$

where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a semi-positive definite function.
Kernel trick: lift x into infinite dimensional feature space by replacing inner products $x^\top x'$ by $k(x, x')$, but never evaluate these features, only the $n \times n$ kernel matrix.

$$\hat{y}' = m(x') = \sum_{i=1}^{n} \alpha_i k(x, x_i)$$

Generally, we don't think about features, we just choose a kernel. But choosing a kernel is implicitly choosing features, and our model only includes functions that are linear combinations of this set of features (the Reproducing Kernel Hilbert Space (RKHS) of k).

Generally, we don't think about features, we just choose a kernel. But choosing a kernel is implicitly choosing features, and our model only includes functions that are linear combinations of this set of features (the Reproducing Kernel Hilbert Space (RKHS) of k).

Example: If (modulo some detail)

$$\phi(x) = (e^{-\frac{(x-c_1)^2}{2\lambda^2}}, \ldots, e^{-\frac{(x-c_N)^2}{2\lambda^2}})$$

then as $N \to \infty$ then

$$\phi(x)^\top \phi(x) = \exp\left(-\frac{(x-x')^2}{2\lambda^2}\right)$$

Generally, we don't think about features, we just choose a kernel. But choosing a kernel is implicitly choosing features, and our model only includes functions that are linear combinations of this set of features (the Reproducing Kernel Hilbert Space (RKHS) of k).

Example: If (modulo some detail)

$$\phi(x) = (e^{-\frac{(x-c_1)^2}{2\lambda^2}}, \ldots, e^{-\frac{(x-c_N)^2}{2\lambda^2}})$$

then as $N \to \infty$ then

$$\phi(x)^\top \phi(x) = \exp\left(-\frac{(x-x')^2}{2\lambda^2}\right)$$

Although our simulator may not lie in the RKHS defined by k, this space is much richer than any parametric regression model (and can be dense in some sets of continuous bounded functions), and is thus more likely to contain an element close to the simulator than any class of models that contains only a finite number of features.

Answer 3: Naturalness of GP framework

Why use **Gaussian** processes as non-parametric models?

Answer 3: Naturalness of GP framework

Why use **Gaussian** processes as non-parametric models?

One answer might come from Bayes linear methods[1].
If we only knew the expectation and variance of some random variables, X and Y, then how should we best do statistics?

[1] statistics without probability

Answer 3: Naturalness of GP framework

Why use **Gaussian** processes as non-parametric models?

One answer might come from Bayes linear methods[1].
If we only knew the expectation and variance of some random variables, X and Y, then how should we best do statistics?

It can been shown, that the best second-order inference we can do to update our beliefs about X given Y is

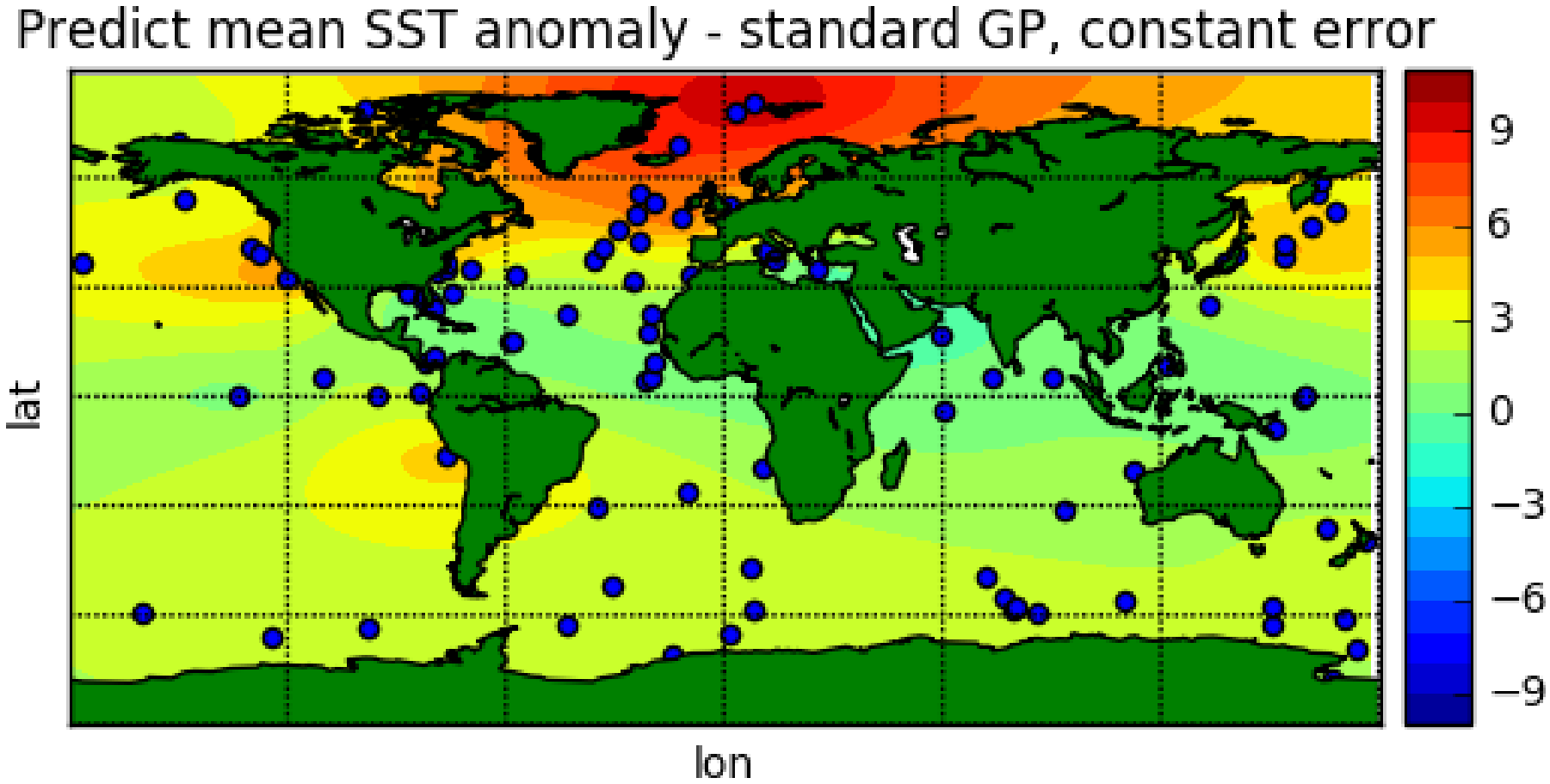
$$\mathbb{E}(X|Y) = \mathbb{E}(X) + \mathbb{C}\text{ov}(X, Y)\mathbb{V}\text{ar}(Y)^{-1}(Y - \mathbb{E}(Y))$$

which is exactly the Gaussian process update for the posterior mean.

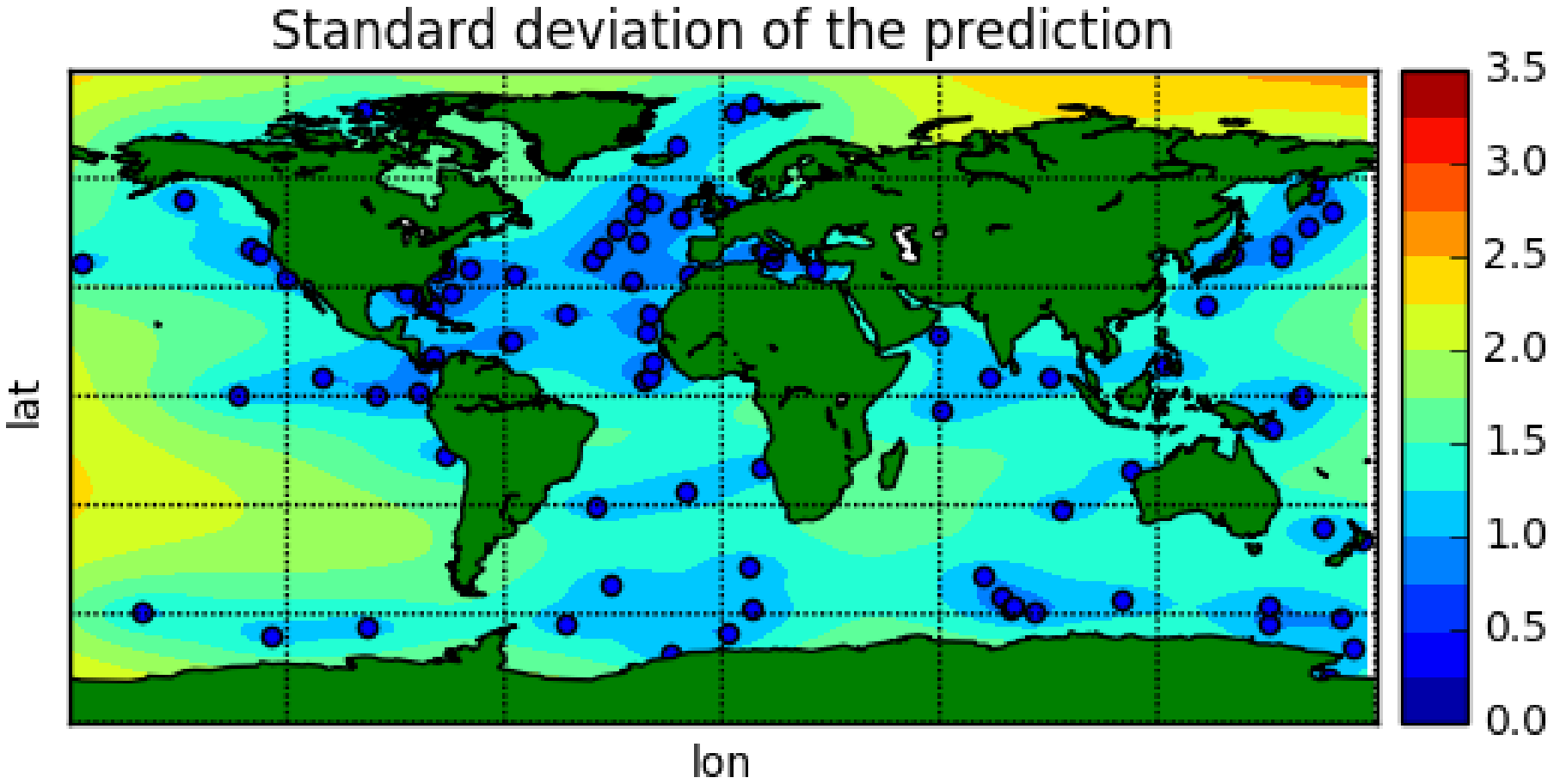
So GPs are in some sense very natural approaches.

[1] statistics without probability

GP prediction



GP prediction - standard deviation



What's wrong?

Interpolation on a sphere

Just giving the GP latitude and longitude coordinates of the observations, means that it doesn't know the data lie on a sphere.

- fix by replacing Euclidean distance with haversine (great circle) distance

Theoretically problematic for GPs,

- only a very small set of covariance functions are positive semi-definite on S^2.
- matters less in practice

Uncertain uncertainty

Heteroscedastic GPs

Observation vector

$$D = (D_1, \ldots, D_n)$$

We're assuming

$$D_i = f(x_i) + \epsilon_i$$

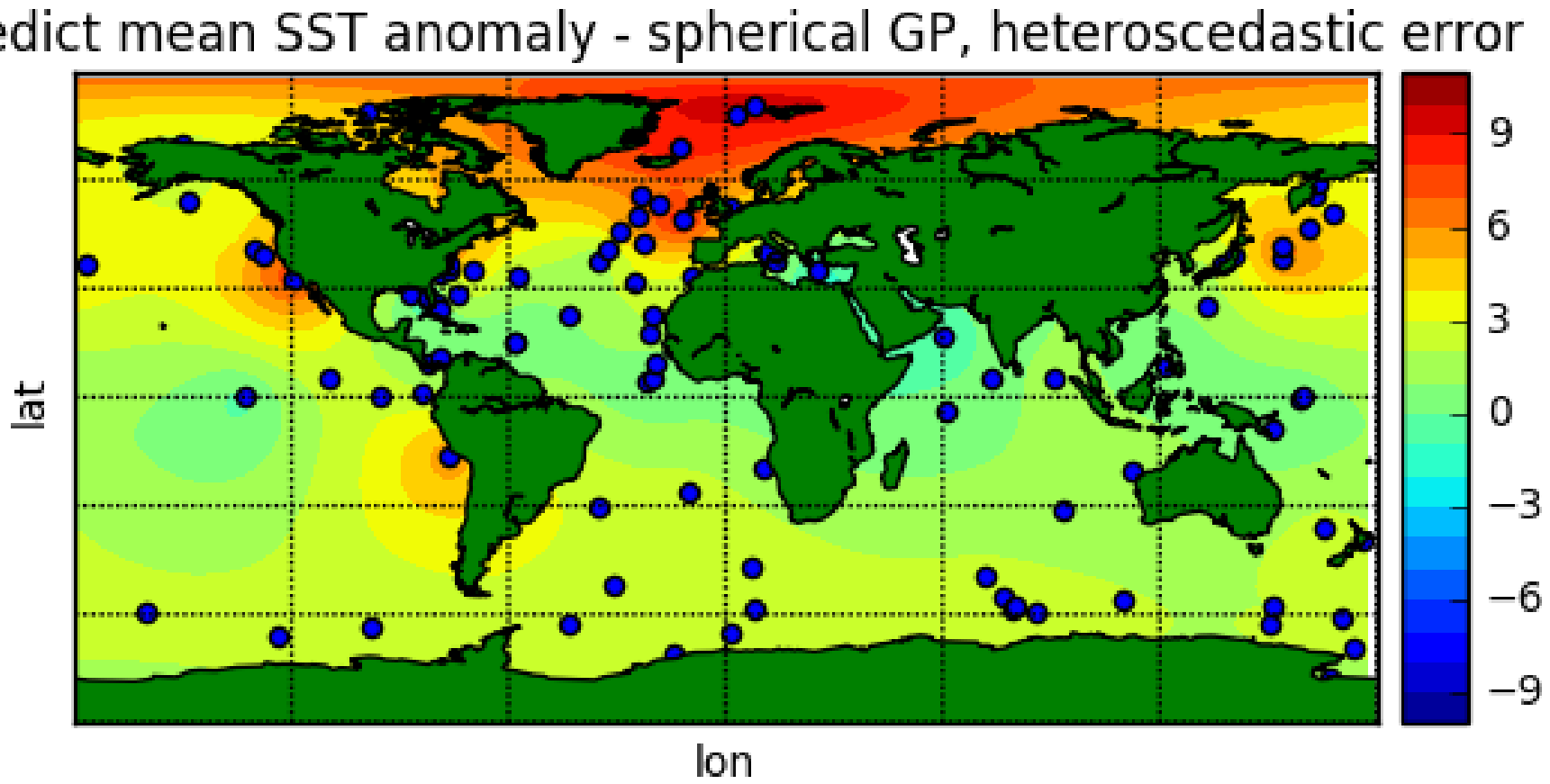
where $f(x)$ is the underlying SST pattern, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top$ is the measurement error.

We use a heteroscedastic measurement model

$$\epsilon \sim N_n(0, V) \text{ where } \mathbb{V}\text{ar}(y) = \tau R$$

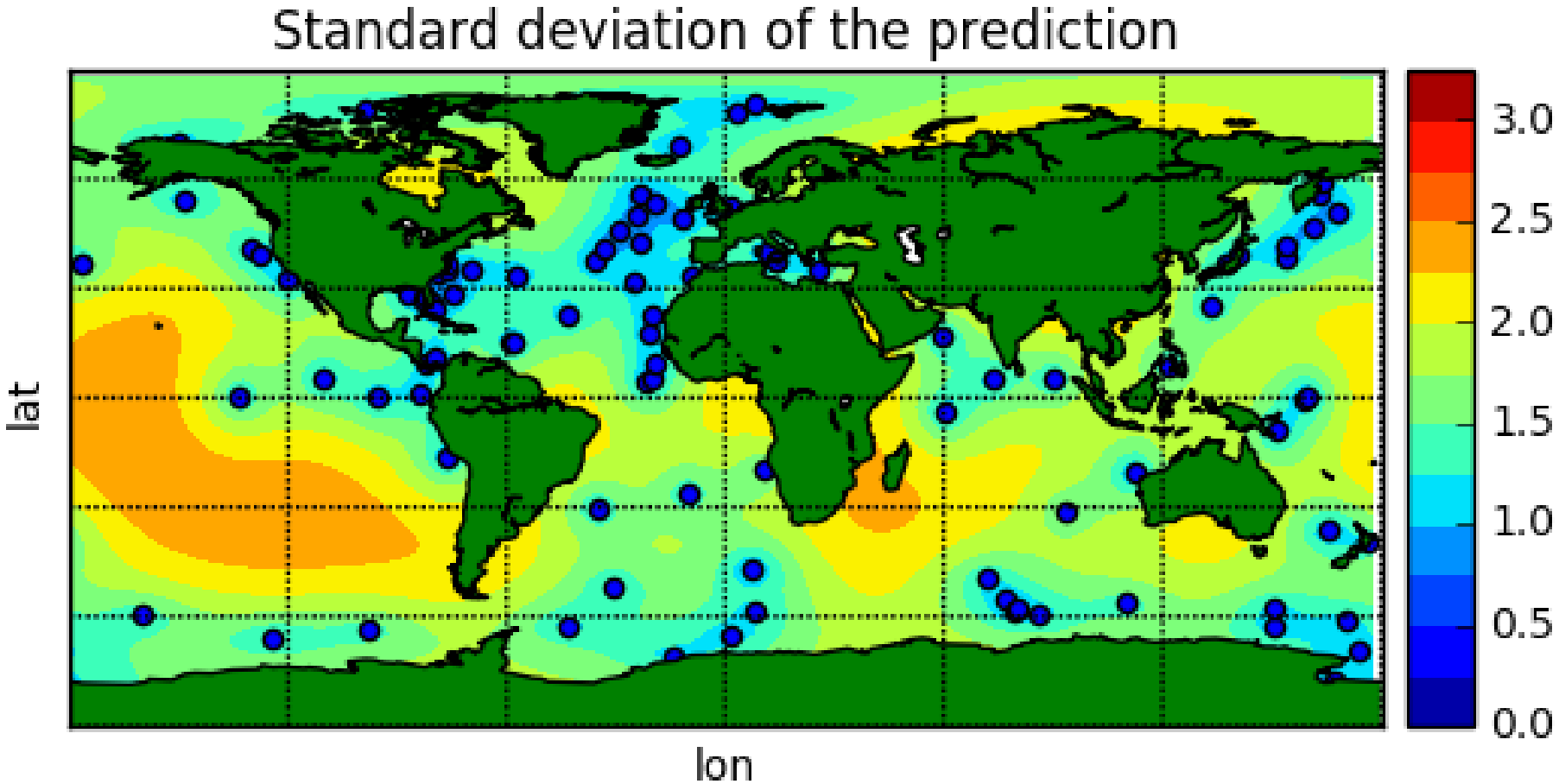
R is a diagonal matrix specifying the ratios of the measurement variances from the qualitative assessment of data quality.
τ is a free parameter scaling the measurement errors which we estimate from the data.

GP prediction



Two observations which are close count less than two far apart (according to a length-scale $\sim 5000km$).
Noisier observations are less influential on the model fit

GP prediction - standard deviation



Code

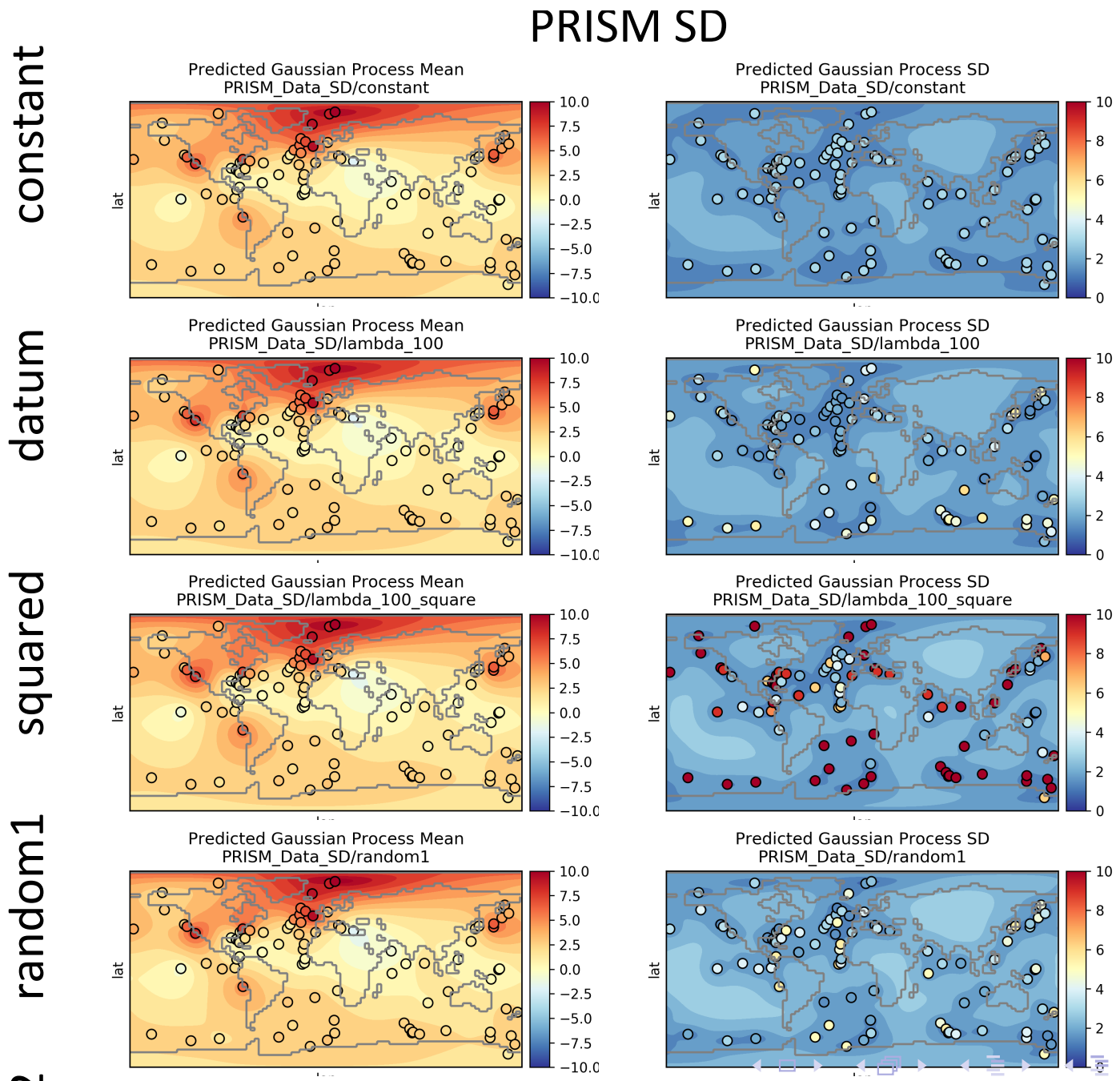
We've added this functionality to GPy.

```
from scaledheteroscedasticgaussian import ScaledHeteroscedasticGaussian
from gp_heteroscedastic_ratios import ScaledHeteroscedasticRegression


k3 = Exponentialhaversine(2, lengthscale=2000)
m3 = ScaledHeteroscedasticRegression(X=X_obs, Y=y_obs, kernel=k3,
                         noise_mult=1., known_variances=var_ratios)
m3.optimize_restarts(10)
```

GMRF requires 100s of lines of code in comparison.

Sensitivity to the uncertainty specification



Scoring

Four scoring functions are tested here:

- LLFC: Loglikelihood using full covariance matrix
- LLVO: Loglikelihood using variance
- RMSEgridded: Root-mean-square error calculated over the whole Gaussian process grid
- RMSEpoints: Root-mean-square error calculated at data locations
  - Default current approach not requiring GP

Doing predictions on the full GCM grid is expensive (30,000 points).

- We can thin it by taking every second value, every third value, etc.

Pseudo data from 405ppm simulation

RMSE - evaluated at data locations

'pseudo-observations' sampled from the GCM simulations at the 95 locations of the PRISM3D dataset using the standard deviation values at each gridpoint calculated across the 30 year climate averaging period as uncertainty estimates.

	scale x thinby	scale	thinby	280 ppm	315 ppm	350 ppm	375 ppm	405 ppm	475 ppm	560 ppm
RMSE - Points	**0.75**	0.75	1	1.46	1.33	1.30	1.25	1.28	1.58	1.98
RMSE - Points	**1**	1	1	1.45	1.32	1.26	1.27	1.28	1.55	1.93
RMSE - Points	**1.5**	0.75	2	1.48	1.34	1.28	1.27	1.29	1.56	1.94
RMSE - Points	**2**	1	2	1.45	1.32	1.31	1.26	1.32	1.60	1.99
RMSE - Points	**2**	2	1	1.48	1.36	1.32	1.30	1.32	1.59	1.96
RMSE - Points	**3**	0.75	4	1.46	1.34	1.29	1.24	1.28	1.56	1.89
RMSE - Points	**4**	1	4	1.50	1.38	1.50	1.38	1.47	1.85	2.25
RMSE - Points	**4**	2	2	1.48	1.34	1.27	1.25	1.26	1.51	1.83
RMSE - Points	**4**	4	1	1.46	1.32	1.25	1.23	1.27	1.54	1.91
RMSE - Points	**8**	2	4	1.50	1.36	1.38	1.31	1.33	1.63	1.98
RMSE - Points	**8**	4	2	1.57	1.38	1.30	1.27	1.24	1.50	1.83
RMSE - Points	**16**	4	4	1.56	1.39	1.27	1.26	1.21	1.41	1.68

Pseudo data from 405ppm simulation

RMSE - gridded across GP prediction

	scale x thinby	scale	thinby	280 ppm	315 ppm	350 ppm	375 ppm	405 ppm	475 ppm	560 ppm
RMSE - Gridded	**0.75**	0.75	1	1.25	1.02	0.85	0.77	0.70	0.86	1.20
RMSE - Gridded	**1**	1	1	1.25	1.02	0.87	0.78	0.73	0.89	1.22
RMSE - Gridded	**1.5**	0.75	2	1.21	0.96	0.80	0.69	0.65	0.84	1.22
RMSE - Gridded	**2**	1	2	1.21	0.96	0.80	0.71	0.67	0.85	1.23
RMSE - Gridded	**2**	2	1	1.22	0.97	0.81	0.70	0.66	0.84	1.21
RMSE - Gridded	**3**	0.75	4	1.21	0.96	0.79	0.69	0.64	0.84	1.22
RMSE - Gridded	**4**	1	4	1.21	0.96	0.80	0.70	0.66	0.85	1.23
RMSE - Gridded	**4**	2	2	1.22	0.97	0.80	0.70	0.67	0.84	1.21
RMSE - Gridded	**4**	4	1	1.20	0.95	0.78	0.68	0.64	0.82	1.20
RMSE - Gridded	**8**	2	4	1.21	0.95	0.78	0.67	0.61	0.79	1.15
RMSE - Gridded	**8**	4	2	1.21	0.94	0.76	0.66	0.59	0.78	1.15
RMSE - Gridded	**16**	4	4	1.25	1.00	0.79	0.71	0.58	0.75	1.13

Pseudo data from 405ppm simulation

Loglikelihood variance only

	scale x thinby	scale	thinby	280 ppm	315 ppm	350 ppm	375 ppm	405 ppm	475 ppm	560 ppm
LL-VO	**0.75**	0.75	1	-120846	-83701	-63621	-54034	-48544	-71898	-127044
LL-VO	**1**	1	1	-71698	-50000	-38500	-32893	-29864	-43999	-76907
LL-VO	**1.5**	0.75	2	-24086	-17390	-14177	-12241	-11512	-16221	-27422
LL-VO	**2**	1	2	-14431	-10351	-8491	-7310	-6876	-9593	-16209
LL-VO	**2**	2	1	-13044	-9443	-7692	-6632	-6226	-8645	-14462
LL-VO	**3**	0.75	4	-6063	-4346	-3511	-3020	-2857	-4086	-6987
LL-VO	**4**	1	4	-3635	-2581	-2143	-1818	-1709	-2478	-4191
LL-VO	**4**	2	2	-3282	-2389	-1956	-1693	-1604	-2206	-3639
LL-VO	**4**	4	1	-3372	-2462	-1993	-1756	-1682	-2301	-3858
LL-VO	**8**	2	4	-749	-549	-449	-398	-378	-519	-854
LL-VO	**8**	4	2	-826	-594	-469	-427	-414	-581	-991
LL-VO	**16**	4	4	-270	-179	-123	-107	-83	-132	-250

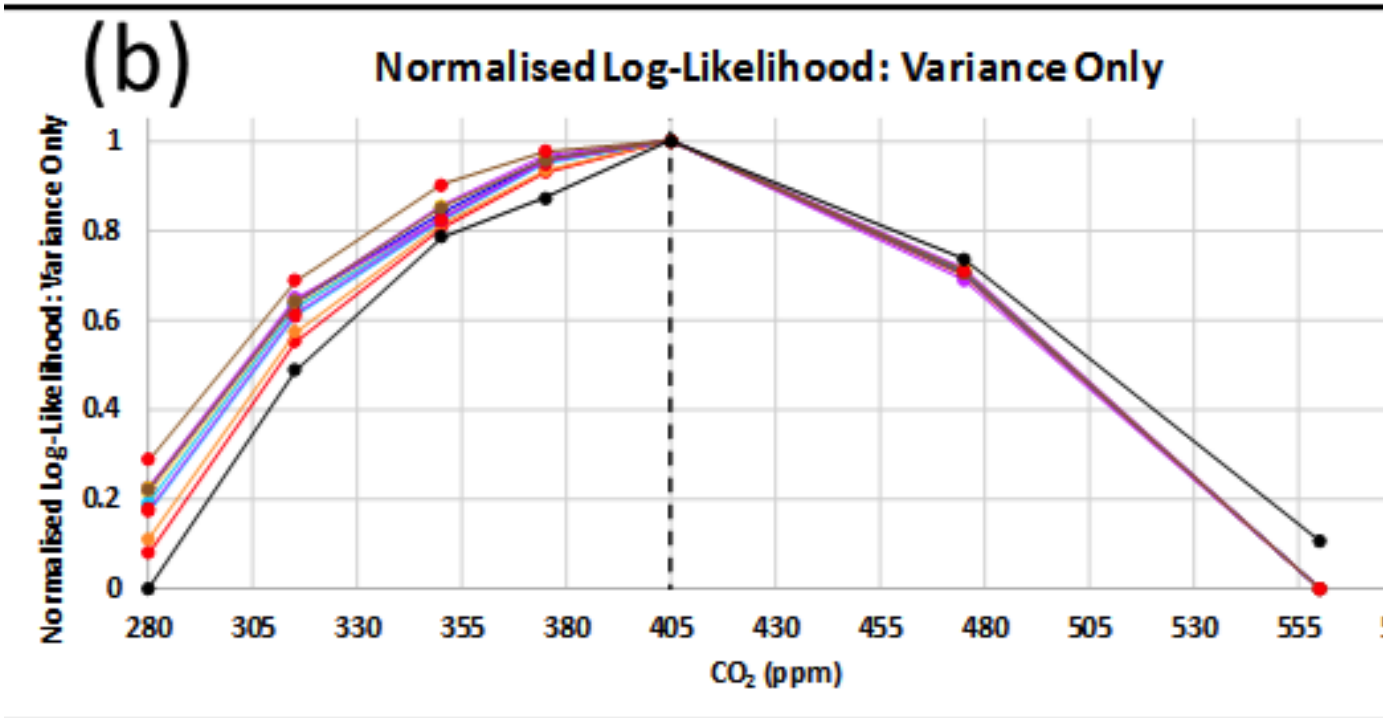
Pseudo data from 405ppm simulation

Loglikelihood full covariance

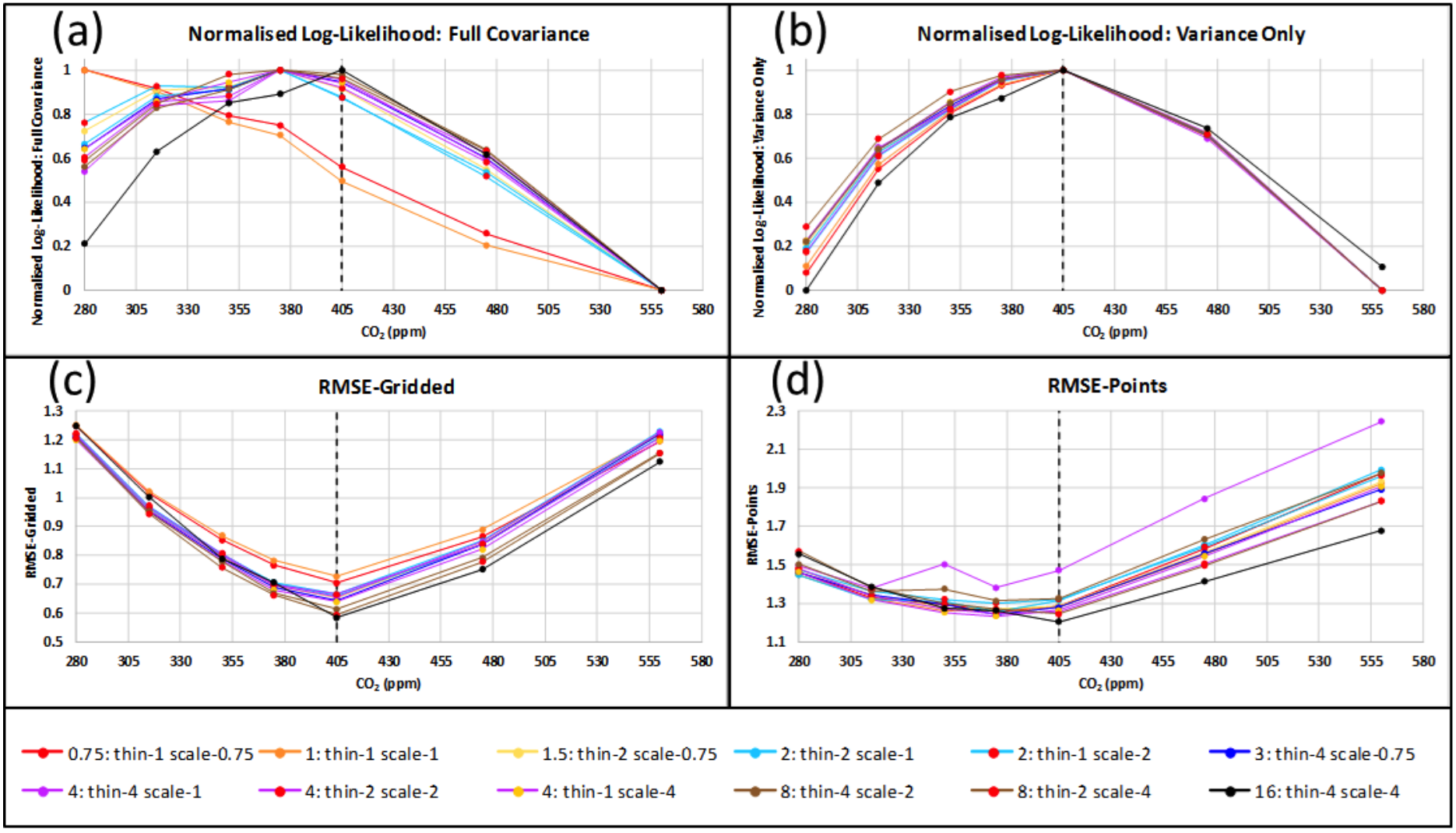
	scale x thinby	scale	thinby	280 ppm	315 ppm	350 ppm	375 ppm	405 ppm	475 ppm	560 ppm
LL-FC	**0.75**	0.75	1	38589	37946	36990	36627	35157	32783	30779
LL-FC	**1**	1	1	9529	8319	6567	5826	3204	-468	-3038
LL-FC	**1.5**	0.75	2	990	1530	1581	1797	1549	476	-1136
LL-FC	**2**	1	2	-1557	-941	-856	-606	-960	-1915	-3432
LL-FC	**2**	2	1	-930	-558	-573	-400	-672	-1477	-2637
LL-FC	**3**	0.75	4	-1123	-793	-725	-600	-676	-1184	-2071
LL-FC	**4**	1	4	-1194	-837	-810	-644	-694	-1123	-1838
LL-FC	**4**	2	2	-935	-704	-680	-572	-646	-956	-1491
LL-FC	**4**	4	1	-886	-689	-614	-563	-614	-906	-1468
LL-FC	**8**	2	4	-393	-307	-279	-251	-258	-368	-575
LL-FC	**8**	4	2	-437	-332	-277	-270	-285	-418	-676
LL-FC	**16**	4	4	-193	-130	-97	-91	-75	-132	-225

Scoring

Log-likelihood variance only



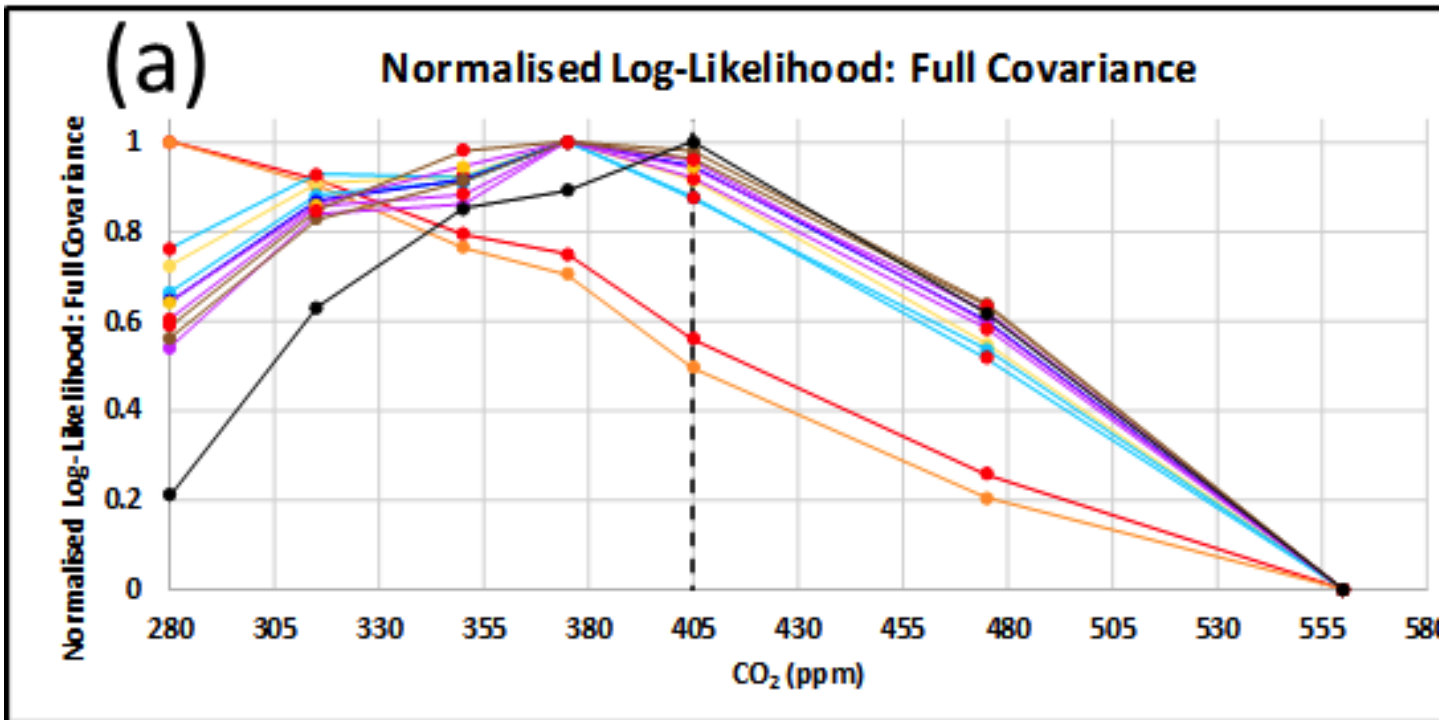
Scoring


(a)
Normalised Log-Likelihood: Full Covariance
Normalised Log-Likelihood: Full Covariance
CO2 (ppm)
(b)
Normalised Log-Likelihood: Variance Only
Normalised Log-Likelihood: Variance Only
CO2 (ppm)
(c)
RMSE-Gridded
RMSE-Gridded
CO2 (ppm)
(d)
RMSE-Points
RMSE-Points
CO2 (ppm)
0.75: thin-1 scale-0.75
1: thin-1 scale-1
1.5: thin-2 scale-0.75
2: thin-2 scale-1
2: thin-1 scale-2
3: thin-4 scale-0.75
4: thin-4 scale-1
4: thin-2 scale-2
4: thin-1 scale-4
8: thin-4 scale-2
8: thin-2 scale-4
16: thin-4 scale-4

Problems

The log-likelihood using the full covariance is unfortunately not stable under subsampling of the GCM grid

- Likely due to numerical problems
- Likelihood also seems to give undue weight to short range correlations



Conclusions and outlook

Gaussian processes rather than Gauss Markov random fields

- Conceptually and computationally simpler, better software
- GPs limit the size of dataset we can consider ($< 10,000$ data points)
- GPs don't allow for complex geometries, e.g., non-spherical earth, continental disruptions to the teleconnection between Pacific and Atlantic etc, but the more flexible GMRF approach would

Proper uncertainty quantification would be better.

- a likelihood function of the data is philosophically backwards
- Model predictions that incorporate uncertainty about what the observations should be under a given scenario would be preferable (forwards models)
- UQ on data is terrible

Code on github (caveat emptor)

More thought on scoring needed

- right feature but wrong place/time, better than no feature

Conclusions and outlook

Gaussian processes rather than Gauss Markov random fields

- Conceptually and computationally simpler, better software
- GPs limit the size of dataset we can consider ($< 10,000$ data points)
- GPs don't allow for complex geometries, e.g., non-spherical earth, continental disruptions to the teleconnection between Pacific and Atlantic etc, but the more flexible GMRF approach would

Proper uncertainty quantification would be better.

- a likelihood function of the data is philosophically backwards
- Model predictions that incorporate uncertainty about what the observations should be under a given scenario would be preferable (forwards models)
- UQ on data is terrible

Code on github (caveat emptor)

More thought on scoring needed

- right feature but wrong place/time, better than no feature

Thank you for listening!