Supplementary material for:

The effects of time-varying observation errors on semi-empirical sea-level projections

Kelsey L. Ruckert[1], Yawen Guan[2], Alexander Bakker[3], Chris E. Forest[1,3,4], Klaus Keller[1,3,5,*]

1 Department of Geosciences, The Pennsylvania State University, University Park, Pennsylvania, USA

2 Department of Statistics, The Pennsylvania State University, University Park, Pennsylvania, USA

3 Earth and Environmental Systems Institute, The Pennsylvania State University, University Park, Pennsylvania, USA

4 Department of Meteorology, The Pennsylvania State University, University Park, Pennsylvania, USA

5 Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

* Email: klaus@psu.edu

1 Detailed description of model-fitting methods

The analyzed methods include a Frequentist Bootstrap and a Bayesian inversion accounting for either heteroskedastic or homoskedastic errors. By using the same sea level and temperature data, we evaluate each method's ability to reconstruct observations with autocorrelated and heteroskedastic data-model residuals.

1.1 Bootstrap method

The Bootstrap method is based on Monte Carlo simulation that provides confidence intervals by resampling from an observed dataset [1]. We adopt the Bootstrap method developed by Solow [1] for correlated observations. If y_t represents sea-level data, model output $f(\theta, t)$, unknown model parameters θ, an unkown residual term R_t, and time t over a hindcast period of 1880-2002, then sea-level observations are modeled as the model output plus an unknown residual term:

$$y_t = f(\theta, t) + R_t. \tag{1}$$

The observations, model outputs, and residuals are vectors from 1 to N, a length of 122 years (Eq. 2).

$$\overrightarrow{y} = \begin{pmatrix} y_{t_1} \\ \vdots \\ y_{t_N} \end{pmatrix}, \overrightarrow{f} = \begin{pmatrix} f(\theta, t_1) \\ \vdots \\ f(\theta, t_N) \end{pmatrix}, \overrightarrow{R} = \begin{pmatrix} \omega_{t_1} + \epsilon_{t_1} \\ \vdots \\ \omega_{t_N} + \epsilon_{t_N} \end{pmatrix}. \tag{2}$$

The residuals are assumed to be a stationary normal AR(1) first-order autoregressive process $\overrightarrow{\omega}$ with an added observation error $\overrightarrow{\epsilon}$,

$$R_t = \omega_t + \epsilon_t. \tag{3}$$

The observation error in the Bootstrap method are assumed to be homoskedastic and independent across time; it is distributed $\sim$ as normal N with zero mean and a common (constant) variance, $\epsilon_t \sim N(0, \sigma_\epsilon^2)$. In contrast, in the heteroskedastic Bayesian method, we substitute in the known variance for ϵ_t (see section SI 1.2). The AR(1) process $\overrightarrow{\omega}$ is characterized by the lag-1 (annual) autoregression coefficient as ρ, the AR(1) innovation variance σ^2_{AR1}, and the white noise as $\overrightarrow{\delta}$. The white noise accounts for model structural uncertainty and internal climate variability. So,

$$\omega_t = \rho \times \omega_{t-1} + \delta_t, \tag{4}$$

with the initial value ω_0 of:

$$\omega_0 \sim N(0, \frac{\sigma^2_{AR1}}{1 - \rho^2}). \tag{5}$$

The white noise is independent and identically distributed (i.i.d.) normal, $\delta_t \sim N(0, \sigma^2_{AR1})$. In the Bootstrap approach, the data-model residuals are defined by subtracting the best model fit $\overrightarrow{f}(\hat{\theta})$ from the observed data:

$$\overrightarrow{R} = \overrightarrow{y} - \overrightarrow{f}(\hat{\theta}). \tag{6}$$

The best model fit $\overrightarrow{f}(\hat{\theta})$ is estimated by minimizing the sum of the absolute residuals using a global optimization method (Differential Evolution) [2]. Using the residuals of the best fit, we estimate the lag-1 (annual) autoregression coefficient ρ with the autocorrelation function in "R". The residuals are then resampled with replacement N times (122 years) to form a Bootstrap sample. Taking the standard deviation of the Bootstrap sample, we estimate the stationary process variance σ^2_ρ. In order to avoid accounting for autocorrelation twice, the stationary process variance is transformed into uncorrelated white noise variance σ^2_{AR1} by taking the square root of the stationary variance times one minus the squared autoregression coefficient:

$$\sigma^2_{AR1} = \sqrt{\sigma^2_\rho (1 - \rho^2)}. \tag{7}$$

Using the uncorrelated white noise variance σ^2_{AR1} and the extracted correlation coefficient ρ, we estimate a new set of Bootstrapped residuals, which preserves the autocorrelation, using Eq. 3, 4, and 5. The resulting Bootstrapped residuals are then added to the best model fit according to Eq. 1, which produces the Bootstrap simulation for the hindcast. We then estimate model parameters from the Bootstrap simulation by minimizing the absolute value of the residuals using Differential Evolution [2]. The fitted parameters are substituted in the model with the temperature forcing RCP8.5 to project a smooth simulation $f(\theta, t)$ of sea level. The smooth simulation is superimposed with the process noise to account for the assumed properties of the data generating process (Eq. 1). The Bootstrap process is repeated 20,000 times to create distributions of parameter estimates and SLR [3].

1.2 Bayesian inversion using Markov chain Monte Carlo (homoskedastic and heteroskedastic)

We implement a Bayesian inversion technique using a Markov chain Monte Carlo (MCMC) algorithm [4–7]. This Bayesian approach uses observed data to update our prior knowledge about model parameters θ to obtain a posterior distribution over a hindcast period from 1880–2002. The overall process follows Bayes' theorem [8], where the probability π of observing the parameters θ given $\mid$ the data $\overrightarrow{y}$ is proportional $\propto$ to the likelihood L of the data given the parameters times the prior probability distribution of the parameters,

$$\underbrace{\pi(\theta \mid \overrightarrow{y})}_{\text{Posterior}} \propto \underbrace{L(\overrightarrow{y} \mid \theta)}_{\text{Likelihood}} \times \underbrace{\pi(\theta)}_{\text{Prior}}. \tag{8}$$

Once the prior distribution and likelihood function is defined, the MCMC algorithm samples from the posterior distribution [4–6].

We construct the likelihood function, assuming that the observational dataset $\overrightarrow{y}$ consists of the model output $\overrightarrow{f}(\theta)$ plus an unknown residual term $\overrightarrow{R}$ (Eq. 1 and 2). The residual term also consists of a stationary normal AR(1) first-order autoregressive process $\overrightarrow{\omega}$ and observation error $\overrightarrow{\epsilon}$ (Eq. 3 and 4). In contrast to the Bootstrap method, this approach accounts for observation error (error assumptions are described below). The uncertain parameters include both the unknown model parameters θ (Table 1) and the unknown parameters σ^2_{AR1} (variance) and ρ (lag-1 autoregressive coefficient). By using an AR(1) likelihood, the method accounts for uncertainty in the autocorrelated process parameters. This autoregressive process contains the combined model structural error, natural variability, and is estimated statistically from the model-data residuals.

To derive the likelihood of the data given the parameters $L(\overrightarrow{y} \mid \Theta)$ we find the data-model residuals as in Eq. 6 and assume they follow a normal distribution $\overrightarrow{R} \sim N(0, \Sigma)$, where Σ is the sum of the variance of AR(1) process $\overrightarrow{\omega}$ and the observation errors $\overrightarrow{\epsilon}$;

$$Var(\overrightarrow{\omega}) = \frac{\sigma^2_{AR1}}{1-\rho^2}\begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{N-1} \\ \rho & 1 & \rho & \cdots & \rho^{N-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{N-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{N-1} & \rho^{N-2} & \rho^{N-3} & \cdots & 1 \end{pmatrix}, Var(\overrightarrow{\epsilon}) = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_3^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_N^2 \end{pmatrix}. \tag{9}$$

$$\Sigma = Var(\overrightarrow{\omega}) + Var(\overrightarrow{\epsilon}) = \frac{\sigma^2_{AR1}}{1-\rho^2}\begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{N-1} \\ \rho & 1 & \rho & \cdots & \rho^{N-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{N-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{N-1} & \rho^{N-2} & \rho^{N-3} & \cdots & 1 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_3^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_N^2 \end{pmatrix}. \tag{10}$$

The likelihood function is then:

$$L(\overrightarrow{y} \mid \Theta) = (\frac{1}{\sqrt{2\pi}})^N \times |\Sigma|^{\frac{-1}{2}} \exp\{\frac{-1}{2}[\overrightarrow{y} - \overrightarrow{f}(\theta)]^{\tau}\Sigma^{-1}[\overrightarrow{y} - \overrightarrow{f}(\theta)]\}. \tag{11}$$

$$\Theta = (\theta, \rho, \sigma_{AR1}). \tag{12}$$

We assess the effects of two different assumptions about observational errors: homoskedastic (homoskedastic Bayesian) or heteroskedastic (heteroskedastic Bayesian). In the homoskedastic Bayesian method, the observation errors are assumed to be homoskedastic, where the errors consist of a normal distribution with zero means and a common variance, $\epsilon_t \sim N(0, \sigma_\epsilon^2)$. This means the variance of the observation errors σ_ϵ^2 is constant through time. In the heteroskedastic Bayesian method, the observational errors are assumed heteroskedastic, where the errors consist of a normal distribution with zero means and a time dependent variance $\sigma_{\epsilon_t}^2$, $\epsilon_t \sim N(0, \sigma_{\epsilon_t}^2)$. The time-dependent variance is given by the measurement errors of the observations provided in Church and White [9].

We adopt uniform priors for the parameters and a gamma distribution prior for the parameter $\frac{1}{\tau}$ (Supplementary Table 1). Our MCMC algorithm uses Metropolis-Hastings updates with a joint normal proposal distribution [4, 5, 7]. We use 5×10^6 and 3×10^6 iterations for analysis. The potential scale reduction factor suggests the Markov chain has converged [10]. We drive the fitted model with RCP8.5 temperatures (obtained from the CMIP5 model output archive: `http://cmip-pcmdi.llnl.gov/cmip5/`) to project smooth simulations $\overrightarrow{f}(\hat{\theta})$ of SLR. Following Eq. 1, the smooth simulations are superimposed with the process noise to account for uncertainty in the autocorrelation and the observation errors. The simulations with process noise represent a probabilistic distribution of SLR and the Markov chains represent the distribution of parameter estimates.

2 Testing hindcast reliability

The reliability of a probabilistic projection is typically defined as the extent in which the forecast captures the actual likelihood of the event being predicted [11]. Simply put, a highly reliable projection is neither over-nor-underconfident. Visually, this is assessed using a reliability diagram that plots the expected fraction for perfect coverage PC versus the fraction of actual coverage AC. The fraction of actual coverage AC is simply:

$$AC = 1 - \left(\frac{\#\text{ of data points outside the credible interval}}{\#\text{ of total data points}}\right). \tag{13}$$

Additionally, we evaluate the skill (measure of performance) of the methods using surprise indices. The surprise index quantifies whether the credible interval covers the appropriate fraction of observations by evaluating the deviation from a perfect 1:1 line [12]. A hindcast that is neither over- nor-underconfident would cover 90% of the data with the 90% credible interval and have a surprise index of zero. We focus on credible intervals from 10 to 100% with added emphasis on evaluating the credible intervals associated with the upper tail-area (i.e., credible intervals far exceeding 90%). The surprise index SI is the difference between the fraction of data covered by the different credible intervals AC minus the expected fraction for perfect coverage PC:

$$SI = AC - PC. \tag{14}$$

In fields such as meteorology the surprise index is often calculated by comparing the forecasts to the real observations. Given the task of deriving multi-decadal projections, we focus on the less challenging test of hindcasts.

3 A perfect model experiment with a known structure

Possible causes for poor performance above the 90% credible level include too few observations, the representation of autocorrelated residuals, the representation of heteroskedastic errors, a large number of parameters, too large or too small prior distributions, and structural model errors. We perform a perfect model experiment with a known structure to analyze the effects of potential causes for the poor performance or underconfident bias in the upper tails (Fig. SI 1, 2, 3). To assess these potential problems, we use a simple model with a known structure:

$$\overrightarrow{y} = \alpha + b\overrightarrow{x} + \overrightarrow{R}, \tag{15}$$

$$\overrightarrow{R} = \overrightarrow{\omega} + \overrightarrow{\epsilon}, \tag{16}$$

where α and b are model parameters, $\overrightarrow{x}$ time, $\overrightarrow{R}$ residuals, and $\overrightarrow{y}$ denotes observations. The residuals consist of either independent and identically distributed (i.i.d.) or AR(1) residuals (generated at random) and the error accounts for either homo-or-heteroskedastic errors. In the first test, we evaluate the effects of autocorrelation representation. We represent autocorrelation in three different ways; as independent and identically distributed (i.e, the residuals are non-correlated), as a known AR(1) value ($\sim$0.37), and as AR(1) with unknown autocorrelation (Fig. SI 1). Each method was run iterations using the same observations generated with $\alpha = 1$, $b = 0.2$, $\overrightarrow{x} = (1, 2, ..., 200)^t$, $\rho = 0.3$, $\sigma = 0.5$, and homoskedastic errors. For the second test, we evaluate the impact of various sample sizes of observations on calibration performance. In this test, we fit the heteroskedastic version of the model to different sample sizes of observations from two to 200 (Fig. SI 2). Each calibration is run 1×10^6 iterations with observations generated with $\alpha = 1$, $b = 0.2$, $\rho = 0.3$, $\sigma = 0.5$, and measurement errors range from 0.07 to 0.02 representing heteroskedastic errors. Last, we fit the three model versions (i.i.d., homoskedastic, and heteroskedastic) to different comparable assumptions about the residuals: (i) observations with i.i.d. residuals assuming non-correlated residuals, (ii) observations with AR(1) residuals assuming homoskedastic errors, and (iii) observations with AR(1) residuals accounting for heteroskedastic errors (measurement error ϵ_t) (Fig. SI 3). The i.i.d. method and the heteroskedastic method are run 1×10^6 iterations while the homoskedastic method is run 5×10^6 iterations. The resulting parameter estimates and projections are compared by their parameter distributions and reliability diagram.

The results of these tests suggest the choice of assumptions can potentially impact model performance. In the simple test case, using the most likely autocorrelation without considering its potential uncertainty may cause an underestimation of the parameter distribution and lead to bias in hindcasts and projections. Neglecting the autocorrelation, i.e., assuming white noise (i.i.d) causes an even larger underestimation (Fig. SI 1). We also find better hindcast reliability diagrams are produced when the stochastic structure of the data is taken into consideration. For example, all methods in the third test perform reasonably well as the i.i.d. data, the AR(1) homoskedastic data, and the AR1 heteroskedastic data are fitted according to their known data structure (Fig. SI 3). More importantly, we find the number of observations greatly impacts the hindcast reliability diagram as relatively low numbers of observations tend to produce underconfident hindcasts especially in the upper tails in the simple test case (Fig. SI 2). Consistent with the results of previous studies [11, 13], these results suggest that too few sea-level observations exist in the reconstruction to resolve low probability outcomes with high reliability using the analyzed methods.

4 Analysis with other semi-empirical sea-level models

4.1 The Rahmstorf (2007) model

4.1.1 Model, data, and method implementation

We test the robustness of our conclusions by running the same analysis with the Rahmstorf [14] sea-level model. This semi-empirical model predicts sea-level rise on an annual time step:

$$\frac{\delta H(t)}{\delta(t)} = \alpha \times [T(t) - T_0], \tag{17}$$

where α is the sensitivity of the rate of sea-level rise to temperature, T is the global-mean surface air temperature, T_0 is the temperature when the sea-level anomaly equals zero, H is the global mean sea level, H_0 is the initial value of sea-level rise (prior bounds based on measurement error [9]. We adopt the same data source as Rahmstorf [14]. We use global mean sea level estimates [9] with respect to the average 1990 sea level, use historical temperature anomalies (with respect to the 20th century) [15] and global-mean surface air temperature anomalies based on the CNRM-CM5 simulation of the RCP 8.5 scenario [16, 17].

As in the main text and described above, we fit the sea-level rise model using three model-fitting methods: a Bootstrap method, a Bayesian method assuming homoskedastic errors, and a Bayesian method accounting for the heteroskedastic nature of the observation errors. We run each method with 2×10^4 (Bootstrap), 1×10^7 (homoskedastic Bayesian), and 2.5×10^7 (heteroskedastic Bayesian) iterations, remove a 1% "burn-in" from the Bayesian method chains, and thin the Bayesian method chains to subsets of 2×10^4. Again, we analysis convergence based on visual inspection and the potential scale reduction factor [6, 10]. All parameters, both model and statistical, are set with a uniform prior distribution (Table SI 1).

4.1.2 Results

The results from running the analysis with the Rahmstorf [14] are consistent with those from running the analysis with the slightly more complex Vermeer and Rahmstorf [18] sea-level rise model. For example, the considered methods produce similar hindcasts and reliability diagrams, but different parameter distributions (Fig. SI 4). The methods predict well at low to intermediate credible intervals, yet perform relatively poorly at the high credible intervals and have a small average surprise index between 1-6% (Fig. SI 4). Additionally, we see that as we account for more known observational properties (i.e., moving from Bootstrap to homoskedastic Bayesian to heteroskedastic Bayesian) the distributions widen, the modes decrease, and the tail areas increase for each parameter (Fig. SI 5 and Table SI 1). More importantly the results also display that the effects of the model-fitting method choice is compounded in the upper tail projections. For instance, the projected sea-level anomaly with a 1% (10^{-2}) probability of being equaled or exceeded in the year 2050 (compared to 1990) varies from 0.35 − 0.43 m (Fig. SI 6e). The heteroskedastic Bayesian method gives a roughly 24% larger sea-level anomaly (associated with the 1% probability) in the year 2050 compared to the sea-level anomaly produced by the Bootstrap method. In 2100, the sea-level anomaly produced by the heteroskedastic Bayesian method with a 1% probability of being equaled or exceeded (1.37 m) differs from the Bootstrap method sea-level anomaly (1.04 m) by 0.33 m; this difference corresponds to a 32% increase (Fig. SI 6f).

4.2 The Grinsted et al. (2010) model

4.2.1 Model, data, and method implementation

Additionally, we test the robustness of our conclusions by running the same analysis with the Ginsted et al. [19] sea-level model:

$$S_{eq} = \alpha \times T + b, \tag{18}$$

$$\frac{\delta S(t)}{\delta(t)} = (S_{eq} - S) \times \frac{1}{\tau}, \tag{19}$$

where α is the sensitivity of the rate of sea-level rise to temperature, T is the global-mean surface air temperature, b is a constant, S_{eq} is the sea-level equilibrium, S is the global mean sea-level, $\frac{1}{\tau}$ is the inverse response times, and t is time. Like the other models we include S_0 (the initial value of sea-level rise) as an uncertain parameter (prior bounds based on measurement error [9]. We use global mean sea level estimates [9] referenced to the 1980-1990 average sea level, use historical temperature anomalies (referenced to the 20th century) [15] for the model-fitting process, and use global-mean surface air temperature anomalies based on the CNRM-CM5 simulation of the RCP 8.5 scenario ([16, 17]; `http://cmip-pcmdi.llnl.gov/cmip5/`) for projections.

We fit the sea-level rise model using three model-fitting methods: a Bootstrap method, a Bayesian method assuming homoskedastic errors, and a Bayesian method accounting for the heteroskedastic nature of the observation errors. We run each method with 2×10^4 (Bootstrap), 5×10^6 (homoskedastic Bayesian), and 3×10^6 (heteroskedastic Bayesian) iterations, remove a 2% "burn-in" from the Bayesian method chains, and thin the Bayesian method chains to subsets of 2×10^4. We set all parameters both model and statistical to a uniform prior distribution with the exception of $\frac{1}{\tau}$, which is set with a gamma distribution prior (Table SI 2).

4.2.2 Results

The results from running this analysis is consistent with running the same analysis with the Rahmstorf [14] and the Vermeer and Rahmstorf [18] sea-level rise model. For example, the considered methods produce similar hindcasts and reliability diagrams where the methods predict well at low to intermediate credible intervals, yet perform relatively poorly at the high credible intervals and have a small average surprise index between 2-4% (Fig. SI 7). Again, we see that as we account for more known observational properties (i.e., moving from Bootstrap to homoskedastic Bayesian to heteroskedastic Bayesian) the distributions widen, the modes shift, and the tail areas increase for each parameter (Fig. SI 8 and Table SI 2). Like the other analyses the results show that the effects are enhanced in the upper tail projections where the projected sea-level anomaly with a 1% (10^{-2}) probability of being equaled or exceeded in the year 2050 varies from 0.35 – 0.43 m (Fig. SI 9e). The heteroskedastic Bayesian method gives a roughly 16% larger sea-level anomaly and a roughly 24% larger sea-level anomaly (associated with the 1% probability) in the year 2050 and 2100 compared to the sea-level anomaly produced by the Bootstrap method. These results display that the main conclusions in the paper are robust between three simple sea-level models.

5 Evaluation of sea-level rise from different temperature scenarios and other analyzes

The considered methods produce visually quite similar median fit hindcasts (Fig. SI 10) and are comparable to other sea-level rise projections in the year 2100 (Fig. SI 11). However, the choice of temperature scenario impacts probabilistic projections, but does not change the main conclusions. To test the robustness of our main conclusions, we investigate the effect of different temperature scenarios for projections. This test is run with the Rahmstorf [14] sea-level model. We project sea-level rise with temperature scenarios created with the global-mean surface air temperature anomalies based on the HadGEM2-ES simulation of the RCP scenarios (RCP8.5, RCP6.0, RCP4.5, and RCP2.6; [16]) as obtained from the CMIP5 model output archive (`http://cmip-pcmdi.llnl.gov/cmip5/`). As in the main text and the results from using the Rahmstorf [14] model, the trend in distribution width is consistent from Bootstrap to homoskedastic Bayesian to heteroskedastic Bayesian (Fig. SI 12). In all cases, the heteroskedastic Bayesian projections produce the largest sea-level anomaly with a 1% (10-2) probability of being equaled or exceeded in the year 2100 (compared to 1990) (Fig. SI 13.

References

[1] Andrew R. Solow. Bootstrapping correlated data. *Journal of the International Association for Mathematical Geology*, 17(7):769–775, 1985.

[2] David Ardia, Kris Boudt, Peter Carl, Katharine Mullen, and Brian G. Peterson. Differential evolution with deoptim: An application to non-convex portfolio optimization (april 5, 2010). *The R Journal*, 3(1):27–34, 2011.

[3] B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1994.

[4] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[5] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[6] W. R. Gilks. *Markov chain Monte Carlo in practice*. Chapman and Hall, London, USA, 1997.

[7] Matti Vihola. Robust adaptive metropolis algorithm with coerced acceptance rate. *Statistics and Computing*, 22(5):997–1008, 2012.

[8] M Bayes and M. Price. An essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, f. r. s. communicated by mr. price, in a letter to john canton, a. m. f. r. s. *Philosophical Transactions*, 53:370–418, 1763.

[9] John A. Church and Neil J. White. A 20th century acceleration in global sea-level rise. *Geophysical Research Letters*, 33(1):n/a–n/a, 2006. L01602.

[10] Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

[11] Michael K. Tippett, Timothy DelSole, and Anthony G. Barnston. Reliability of regression-corrected climate forecasts. *Journal of Climate*, 27(9):3393–3404, 2014.

[12] Chiara Scotti. Surprise and uncertainty indexes: Real-time aggregation of real-activity macro-surprises. *Journal of Monetary Economics*, 82:1 – 19, 2016.

[13] David S. Richardson. Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quarterly Journal of the Royal Meteorological Society*, 127(577):2473–2489, 2001.

[14] Stefan Rahmstorf. A semi-empirical approach to projecting future sea-level rise. *Science*, 315(5810):368–370, 2007.

[15] Thomas M. Smith, Richard W. Reynolds, Thomas C.Peterson, and Jay Lawrimore. Improvements to noaa's historical merged land–ocean surface temperature analysis (1880–2006). *Journal of Climate*, 21(10):2283–2296, 2008.

[16] Malte Meinshausen, S. J. Smith, K. Calvin, J. S. Daniel, M. L. T. Kainuma, J-F. Lamarque, K. Matsumoto, S. A. Montzka, S. C. B. Raper, K. Riahi, A. Thomson, G. J. M. Velders, and D.P. P. van Vuuren. The rcp greenhouse gas concentrations and their extensions from 1765 to 2300. *Climatic Change*, 109(1):213–241, 2011.

[17] Keywan Riahi, Shilpa Rao, Volker Krey, Cheolhung Cho, Vadim Chirkov, Guenther Fischer, Georg Kindermann, Nebojsa Nakicenovic, and Peter Rafaj. Rcp 8.5—a scenario of comparatively high greenhouse gas emissions. *Climatic Change*, 109(1):33–57, 2011.

[18] Martin Vermeer and Stefan Rahmstorf. Global sea level linked to global temperature. *Proceedings of the National Academy of Sciences*, 106(51):21527–21532, 2009.

[19] Aslak Grinsted, J. C. Moore, and S. Jevrejeva. Reconstructing sea level from paleo and projected temperatures 200 to 2100 ad. *Climate Dynamics*, 34(4):461–472, 2010.

[20] Stefan Rahmstorf, Mahé Perrette, and Martin Vermeer. Testing the robustness of semi-empirical sea level projections. *Climate Dynamics*, 39(3):861–875, 2012.

[21] Robert E. Kopp, Radley M. Horton, Christopher M. Little, Jerry X. Mitrovica, Michael Oppenheimer, D. J. Rasmussen, Benjamin H. Strauss, and Claudia Tebaldi. Probabilistic 21st and 22nd century sea-level projections at a global network of tide-gauge sites. *Earth's Future*, 2(8):383–406, 2014. 2014EF000239.

[22] Robert E. Kopp, Andrew C. Kemp, Klaus Bittermann, Benjamin P. Horton, Jeffrey P. Donnelly, W. Roland Gehrels, Carling C. Hay, Jerry X. Mitrovica, Eric D. Morrow, and Stefan Rahmstorf. Temperature-driven global sea-level variability in the common era. *Proceedings of the National Academy of Sciences*, 113(11):E1434–E1441, 2016.

Table 1: **Prior uniform distributions and fitted median, mode, and 99% estimates for the parameters using the Rahmstorf [14] model.**

Parameter	Sym.	Units	Prior min.	Prior max.		Bootstrap	Homoskedastic Bayesian	Heteroskedastic Bayesian
Sea-level rise sensitivity	α	cm/yr/C	0	2	Med.	0.3	0.27	0.2
					Mode	0.32	0.28	0.16
					99%	0.39	0.4	0.55
Equilibrium temp.	T_0	C	-3	2	Med.	-0.56	-0.61	-0.79
					Mode	-0.52	-0.57	-0.62
					99%	-0.44	-0.42	-0.3
Initial sea-level anomaly	H_0	cm	-16.8	-12.5	Med.	-15.1	-15.3	-15.4
					Mode	-14.9	-15.3	-16.6
					99%	-14.4	-13.9	-12.6
Lag-1 autoregression coefficient	ρ	-	-0.99	0.99	Med.	-	0.73	0.82
					Mode	0.67	0.7	0.99
					99%	-	0.93	0.99
White noise standard error	σ_{AR1}	cm	0	1	Med.	0.48	0.49	*
					Mode	0.43	0.49	*
					99%	0.54	0.58	*

*The standard error is not depicted because observational error is separate from the model error, whereas the standard error in the Bootstrap and homoskedastic Bayesian methods defines model and observational error as one term

Table 2: **Prior distributions and fitted median, mode, and 99% estimates for the parameters using the Grinsted et al. [19] model.**

Parameter	Sym.	Units	Prior min.	Prior max.		Bootstrap	Homoskedastic Bayesian	Heteroskedastic Bayesian
Sea-level rise sensitivity	α	m/yr/C	0	4	Med.	0.5	2.0	1.8
					Mode	0.5	1.5	0.79
					99%	2.6	3.9	3.9
Constant	b	m	-1	2.5	Med.	0.08	1.0	1.1
					Mode	-0.05	0.6	0.6
					99%	1.3	2.3	2.4
Initial sea-level anomaly	S_0	m	-0.168	-0.125	Med.	-0.158	-0.156	-0.156
					Mode	-0.157	-0.156	-0.163
					99%	-0.149	-0.144	-0.129
Inverse response time	$\frac{1}{\tau}$	1/years	0	1	Med.	0.0116	0.0016	0.0014
					Mode	0.0142	0.001	0.001
					99%	0.0245	0.0056	0.006
Lag-1 autoregression coefficient	ρ	-	-0.99	0.99	Med.	-	0.70	0.81
					Mode	0.60	0.70	0.99
					99%	-	0.90	0.99
White noise standard error	σ_{AR1}	m	0	1	Med.	0.0048	0.0049	*
					Mode	0.0046	0.005	*
					99%	0.0054	0.0058	*

*The standard error is not depicted because observational error is separate from the model error, whereas the standard error in the Bootstrap and homoskedastic Bayesian methods defines model and observational error as one term

Table 3: **List of climate modeling groups that participated in CMIP and whose model output we used in this study (archived at `http://cmip-pcmdi.llnl.gov/cmip5/`).** This table is modified from `http://cmip-pcmdi.llnl.gov/cmip5/docs/CMIP5_modeling_groups.docx`, accessed 13 June 2016

Modeling Center (or Group)	Institute ID	Model Name
Centre National de Recherches Météorologiques / Centre Européen de Recherche et Formation Avancée en Calcul Scientifique	CRM-CERFACS	CNRM-CM5
Met Office Hadley Centre (additional HadGEM2-ES realizations contributed by Instituto Nacional de Pesquisas Espaciais)	MOHC (additional realizations by INPE)	HadCEM2-ES

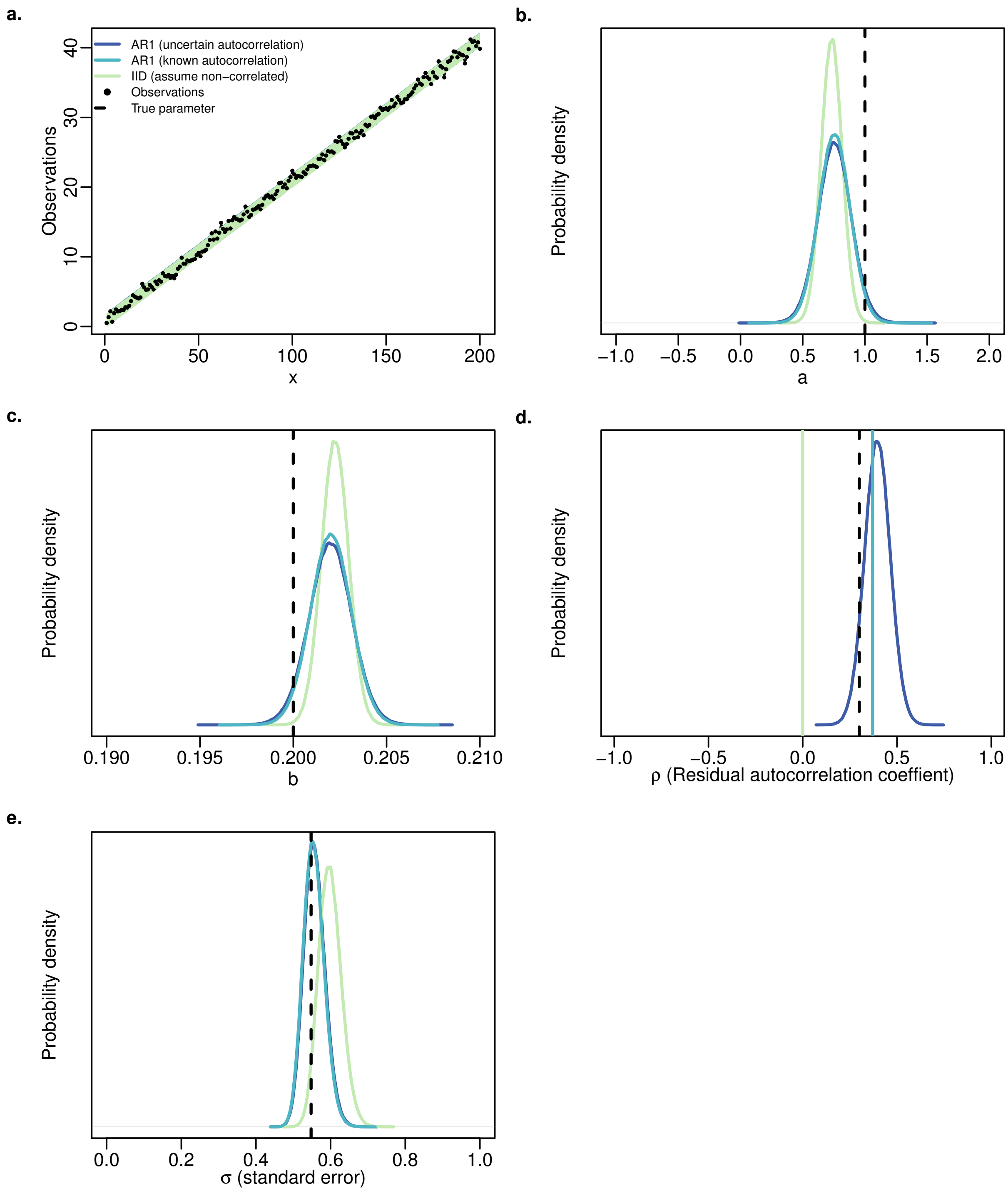

Figure 1: **Test of the impact of fitting a model to observations with autocorrelated residuals with and without accounting for autocorrelation.** Shown are the 90% credible interval (a) projections and (b–d) parameter distributions for accounting uncertain autocorrelation (blue), for accounting for known autocorrelation (turquoise), and for assuming non-correlated data (green). The points represent 200 observations with autocorrelated residuals. The dashed lines in panels b through d represent the true parameter value used to create the 200 observations.

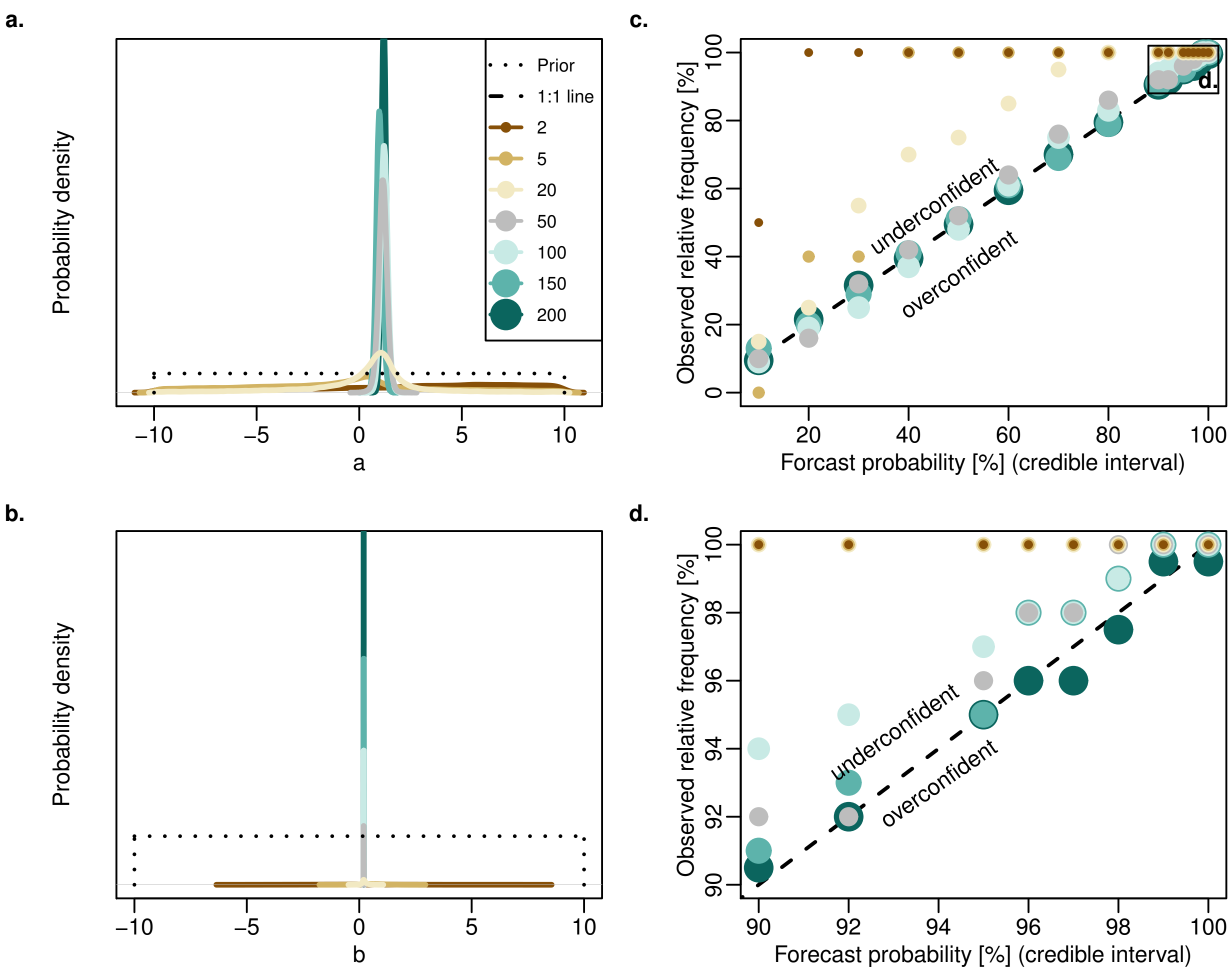

Figure 2: **Test of the impact of increasing the number of observations on the (a-b) probability density function and (c-d) reliability diagram in the simple linear test case for autocorrelated and heteroskedastic residuals.** The dashed lines in panel a and b represent the prior distribution used in the Bayesian inversion, whereas in panel c and d it represents the perfect model outcome. Panel c shows the reliability diagram from 10 to 100% with a zoomed in view of 90-100% in panel d. The solid lines and points represent different numbers of observations (2-200) used in model-fitting.

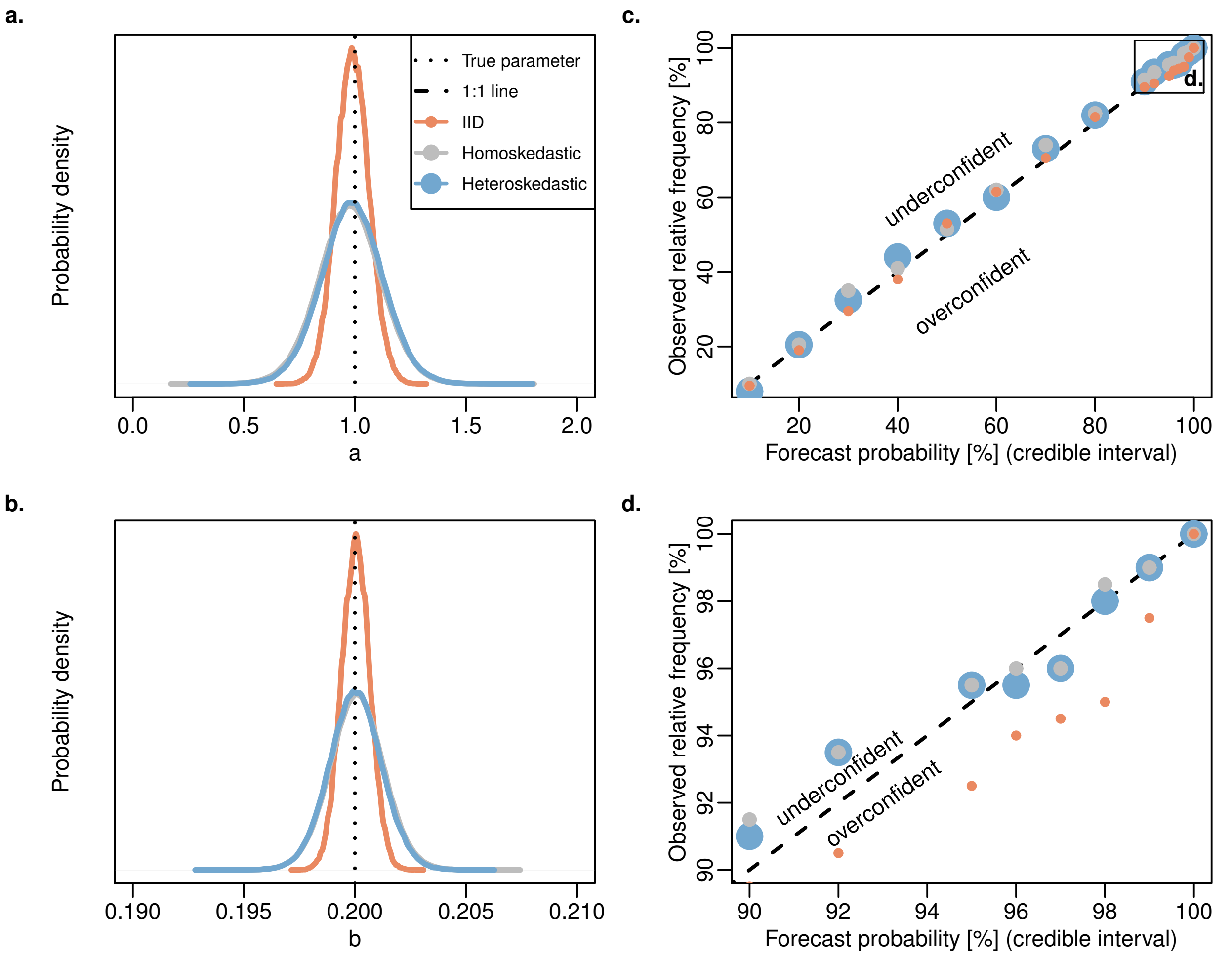

Figure 3: **Test of the impact of varying assumptions about the observation properties on the (a-b) probability density function and (c-d) reliability diagram in the simple linear test case.** The orange lines and points represent the output from fitting 200 i.i.d. residuals assuming i.i.d. residuals. The gray lines and points represent the output from fitting 200 autocorrelated residuals assuming autocorrelated residuals. The blue lines and points represent the output from fitting 200 autocorrelated and heteroskedastic residuals assuming autocorrelation and heteroskedasticity. The dashed lines in panels a and b represent the true parameter value used to create the 200 observations, whereas in panel c and d it represents the perfect model outcome. Panel c shows the reliability diagram from 10 to 100% with a zoomed in view of 90-100% in panel d

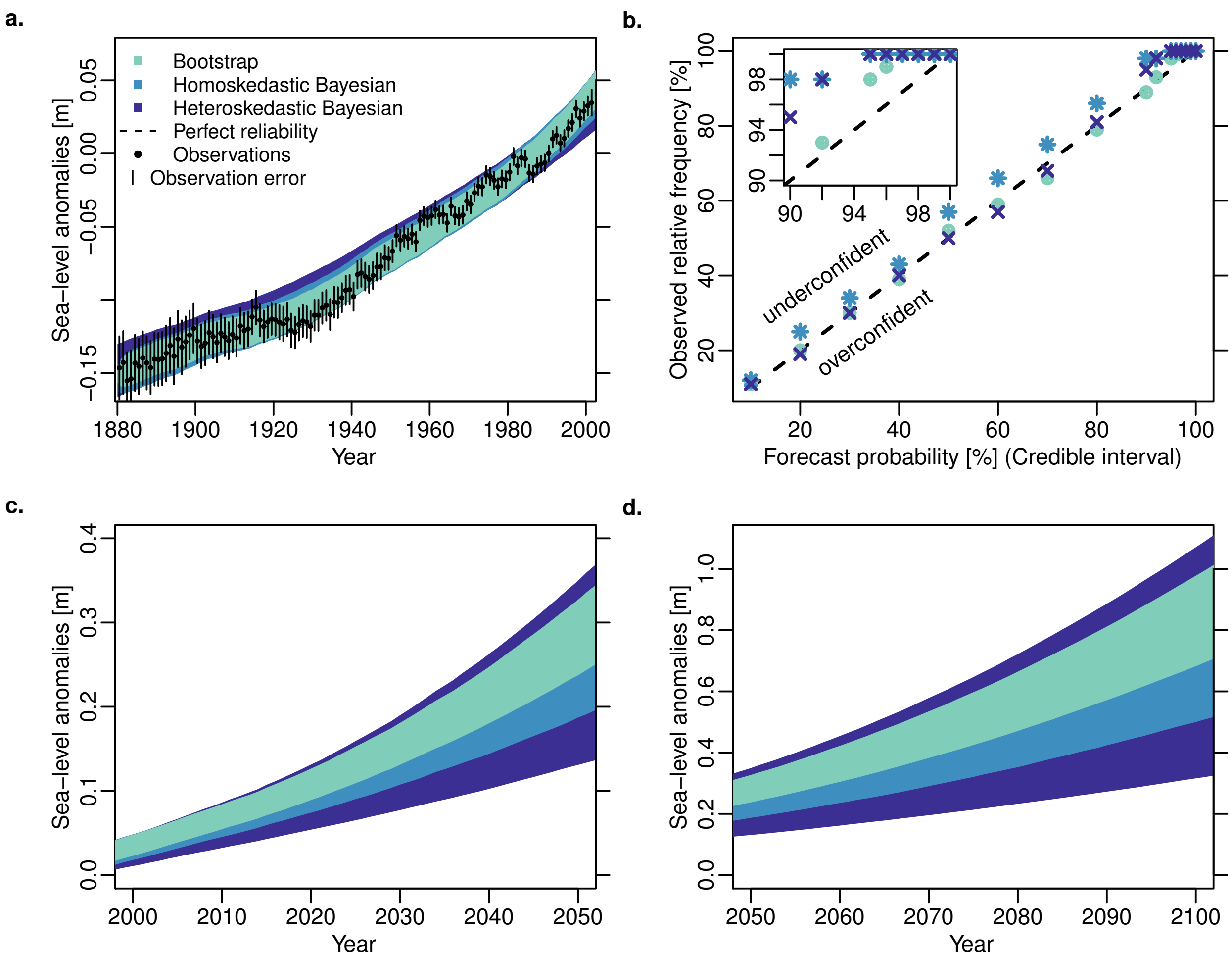

Figure 4: **Comparison of sea-level rise hindcasts and projections and the reliability diagram generated using the Rahmstorf [14] model.** Panel a, c, and d display the 90% credible interval for each method along with the synthesized observations (points) and their associated measurement error. The color ramp from light green to dark blue represents accounting for more known observational properties (Bootstrap to homoskedastic Bayesian to heteroskedastic Bayesian). The reliability diagram in panel b analyzes the hindcast credible intervals produced from each method from 10 to 100% in increments of 10, whereas the subplot in panel b zooms in on the credible intervals from 90 to 100%.

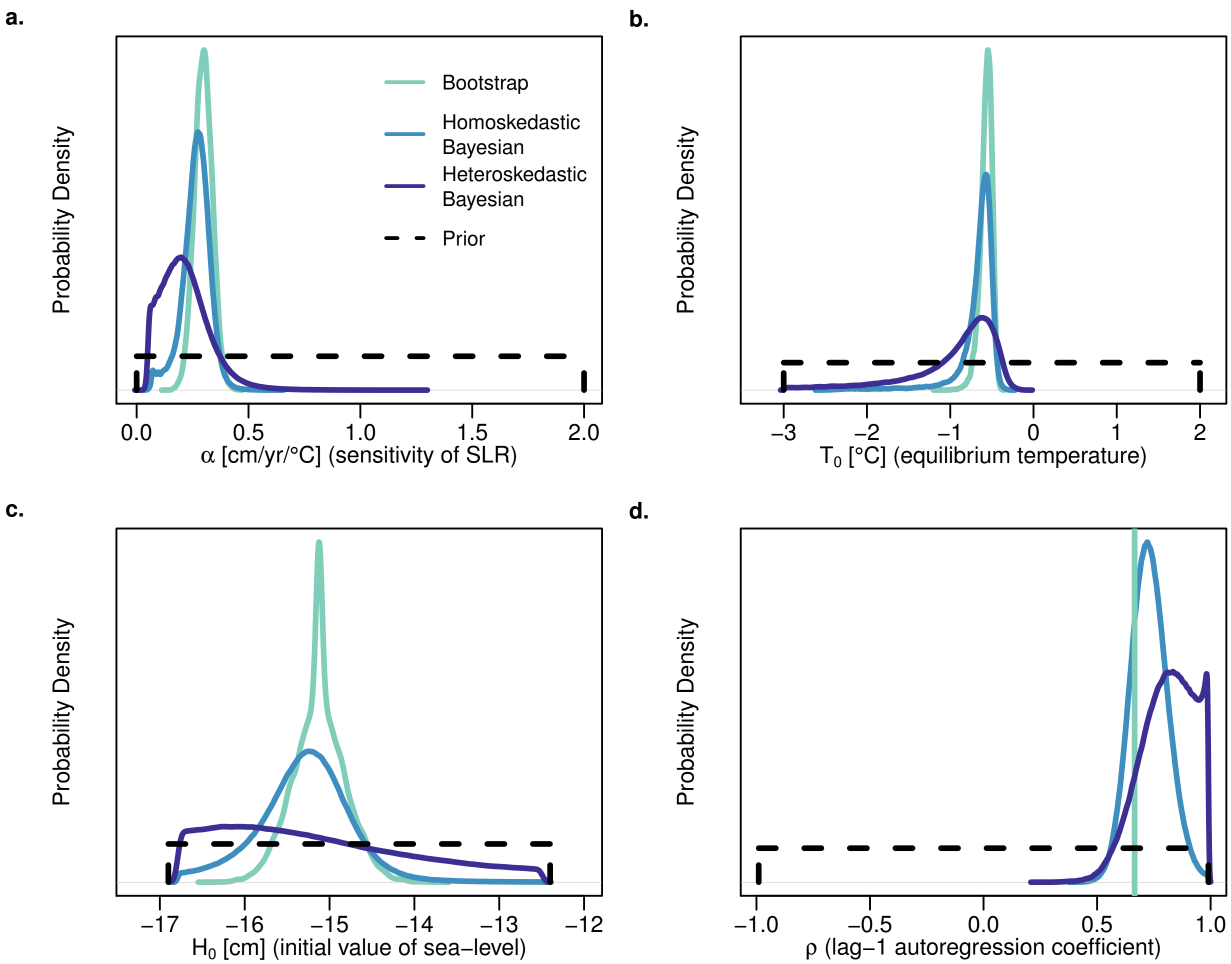

Figure 5: **Marginal probability density functions of the estimated model [14] and statistical parameters.** Shown are (panel a, α) sensitivity of sea-level to changes in temperature, (panel b, T_0) temperature when the sea-level anomaly is zero, (panel c, H_0) initial sea-level anomaly in the year 1880, and (panel d, ρ) the autoregression coefficient. The dashed lines are the prior parameter distribution used in the Bayesian methods.

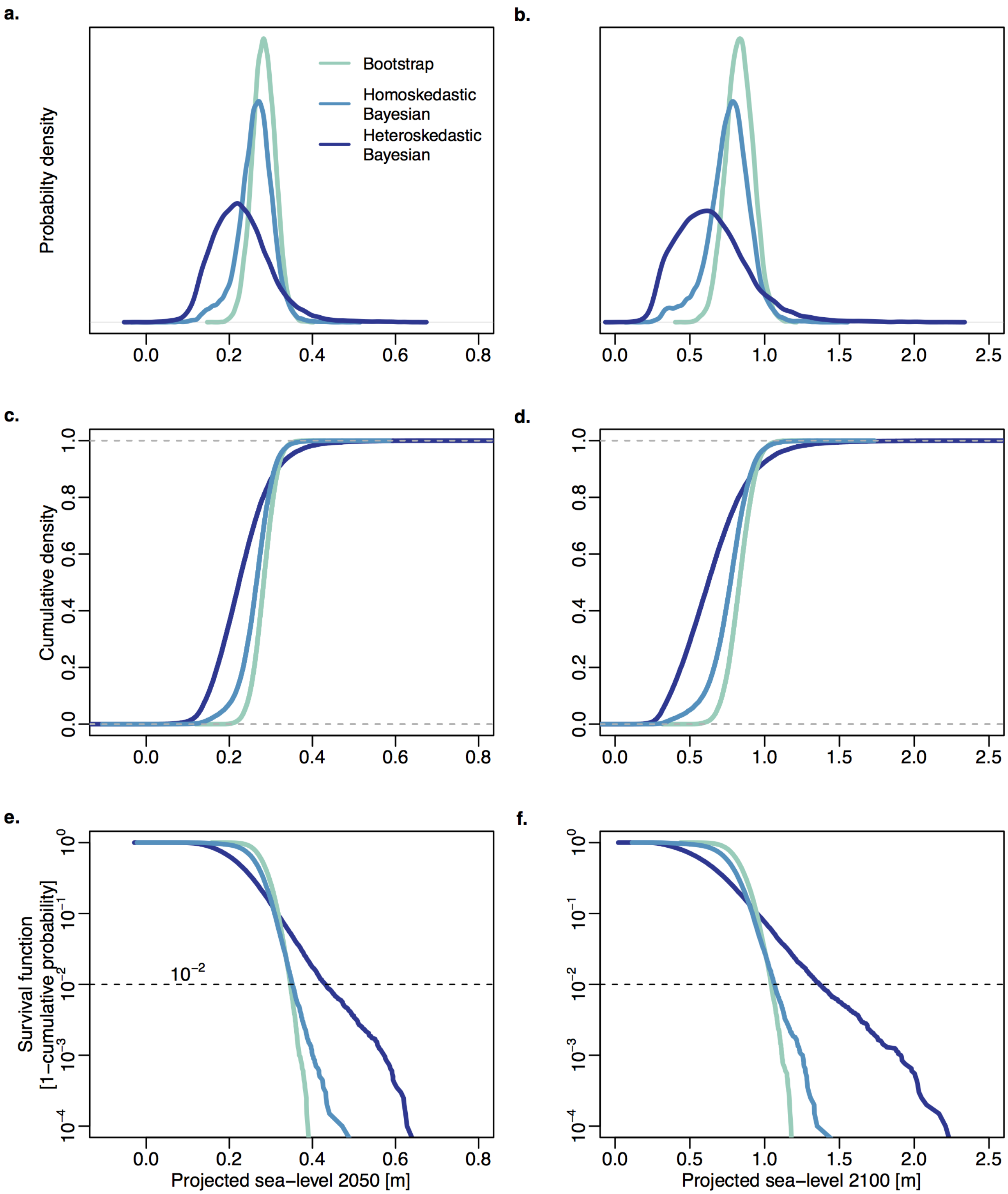

Figure 6: **Projections for global mean sea-level rise using the Rahmstorf [14] model in (a, c, and e) 2050 and (b, d, and f) 2100 determined for the different methods presented as the (a-b) probability density function, the(c-d) cumulative density function, and the (e-f) survival function.** The horizontal dashed lines in the survival function represent the 1% (10^{-2}) probability.

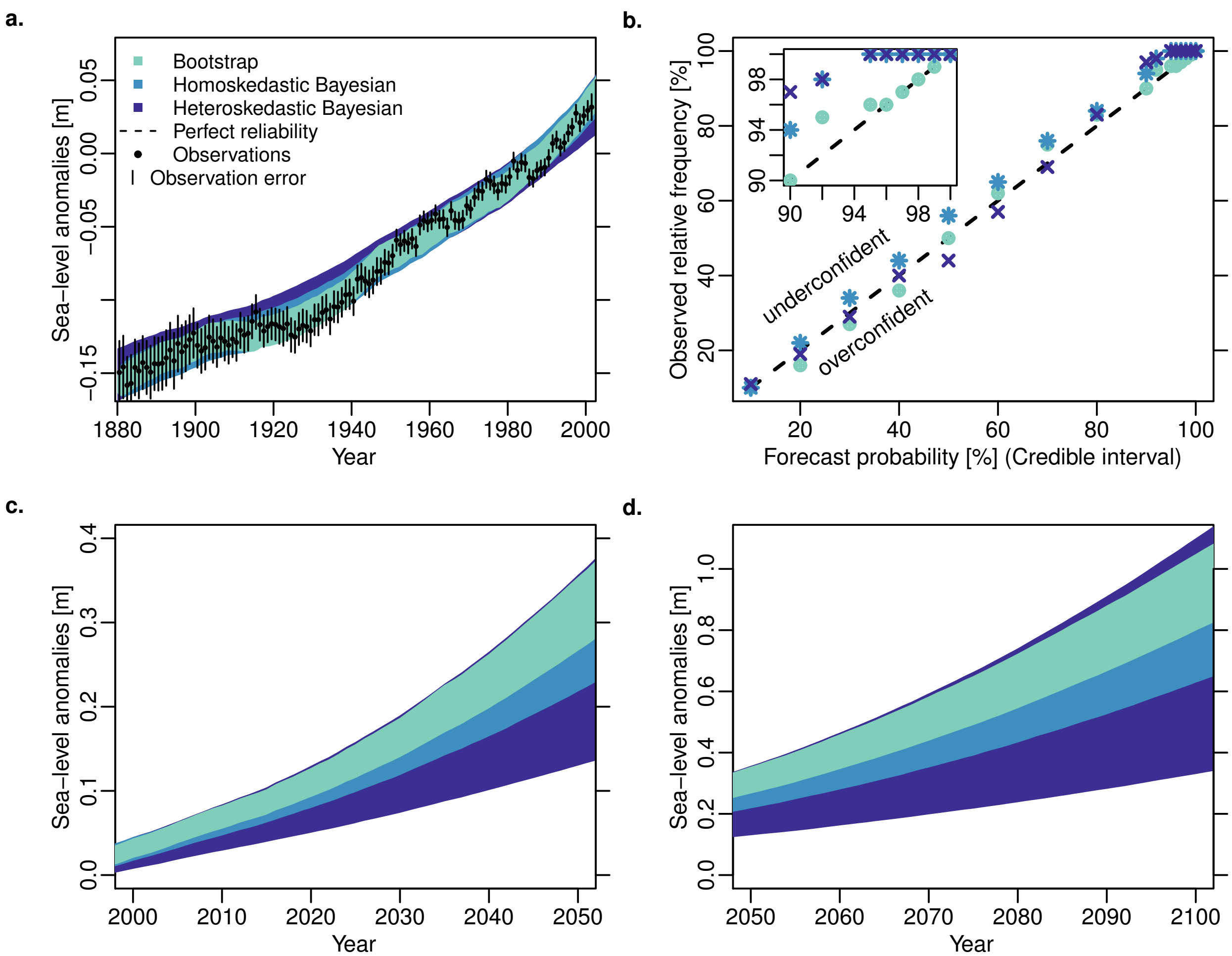

Figure 7: **Comparison of sea-level rise hindcasts and projections and the reliability diagram generated using the Grinsted et al. [19] model.** Panel a, c, and d display the 90% credible interval for each method along with the synthesized observations (points) and their associated measurement error. The color ramp from light green to dark blue represents accounting for more known observational properties (Bootstrap to homoskedastic Bayesian to heteroskedastic Bayesian). The reliability diagram in panel b analyzes the hindcast credible intervals produced from each method from 10 to 100% in increments of 10, whereas the subplot in panel b zooms in on the credible intervals from 90 to 100%.

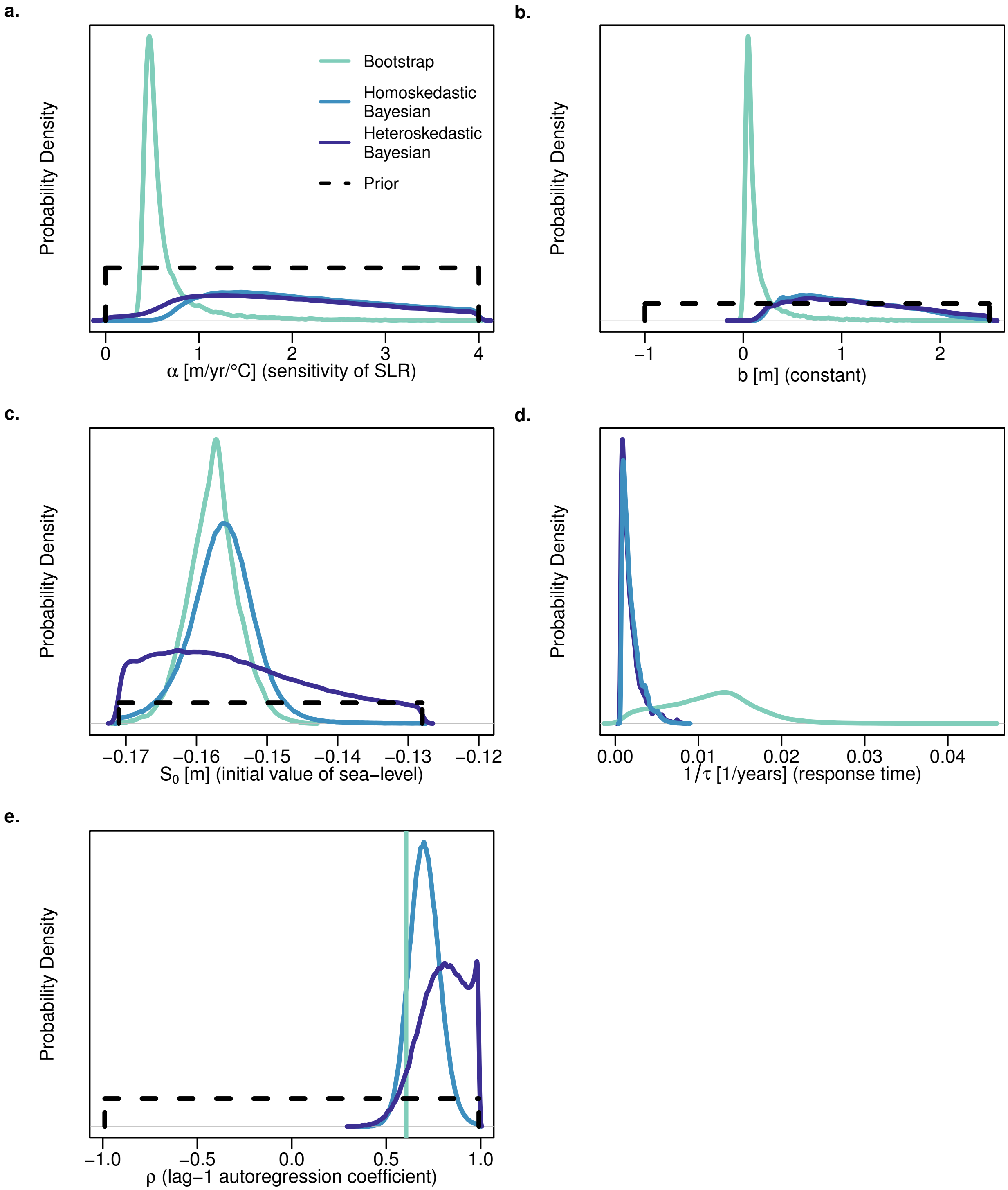

Figure 8: **Marginal probability density functions of the estimated model [19] and statistical parameters.** Shown are (panel a, α) sensitivity of sea-level to changes in temperature, (b, panel b)a constant, (panel c, S_0) initial sea-level anomaly in the year 1880, and (panel d, $\frac{1}{\tau}$) the inverse of the response time, and (panel e, ρ) the autoregression coefficient. The dashed lines are the prior parameter distribution used in the Bayesian methods. The prior for $\frac{1}{\tau}$ is not shown.

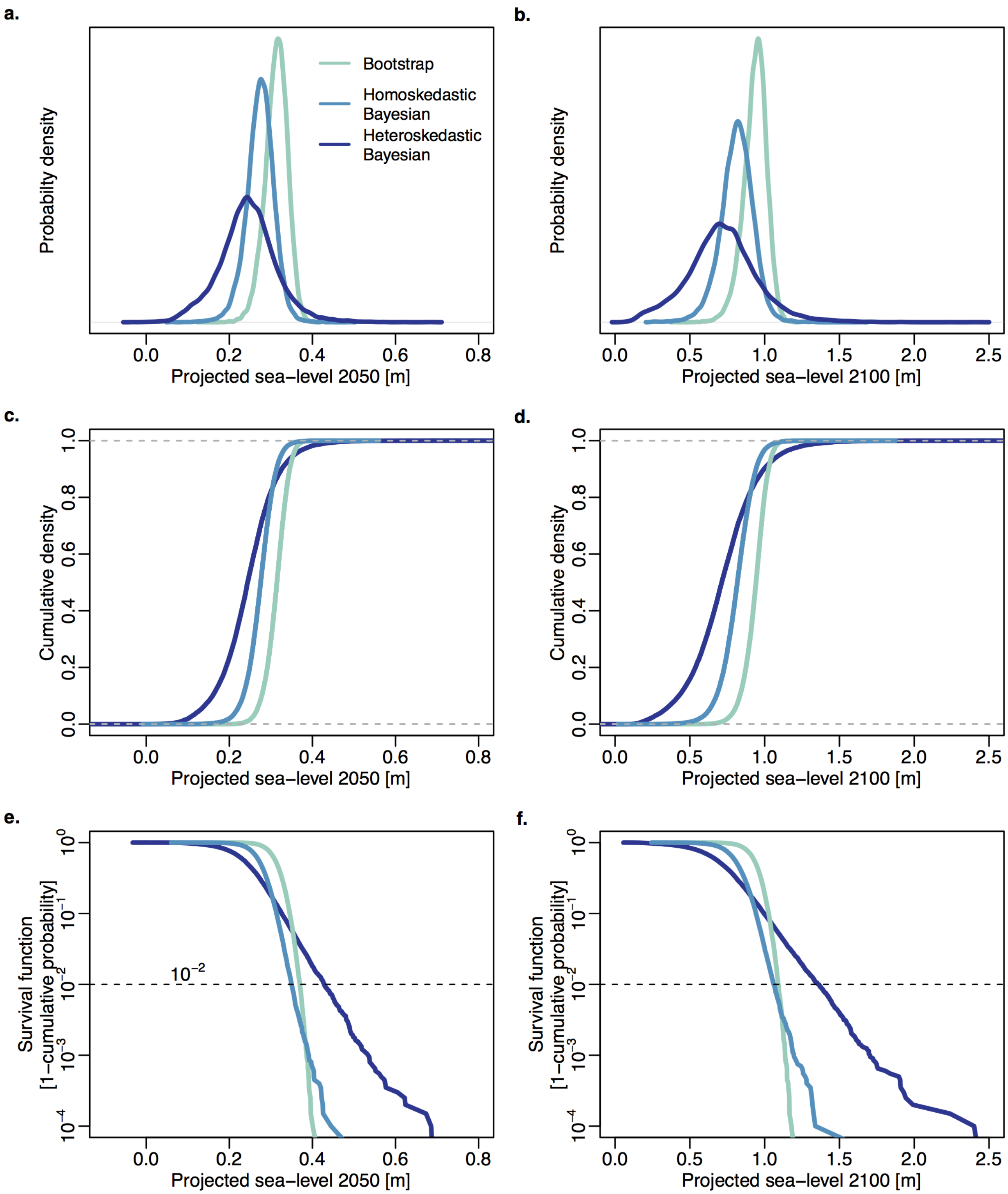

Figure 9: **Projections for global mean sea-level rise using the Grinsted et al. [19] model in (a, c, and e) 2050 and (b, d, and f) 2100 determined for the different methods presented as the (a-b) probability density function, the(c-d) cumulative density function, and the (e-f) survival function.** The horizontal dashed lines in the survival function represent the 1% (10^{-2}) probability.

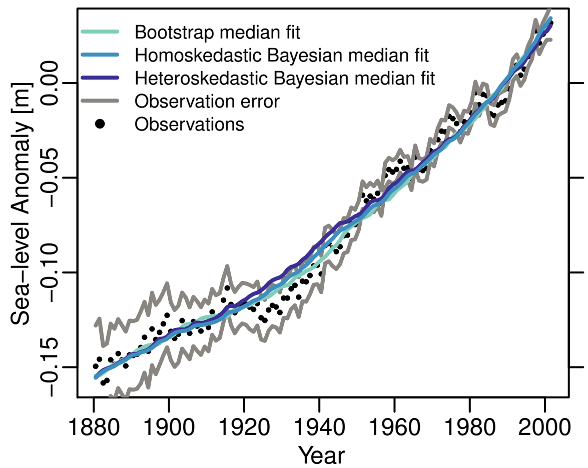

Figure 10: **Median fit hindcasts of global mean sea-level observations of the considered methods.** The median fit for the methods are generated from the median estimate of the model parameters. The gray lines represent the observational error in the sea-level record derived from tide gauge measurements and the points show the synthesized observations in relation to the average 1980-2000 global mean sea level [9].

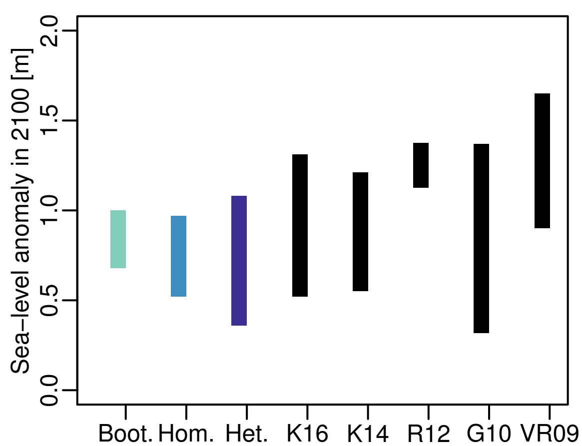

Figure 11: **Projected range of global mean sea-level for the year 2100 by several studies. Each study uses a RCP8.5 temperature scenario or the equivalent SRES A2 temperature scenario.** The green, blue, and navy bars correspond to the 5-95% projections from Bootstrap, homoskedastic Bayesian, and heteroskedastic Bayesian. The black bars represent a few 5-95% projections described in other sea-level analyses (i.e., Vermeer and Rahmstorf [18] (VR09), Grinsted et al. [19] (G10), Rahmstorf et al. [20] (R12), Kopp et al. [21] (K14), and Kopp et al. [22] (K16)). Each of these studies differ by model, methods, assumptions, and data.

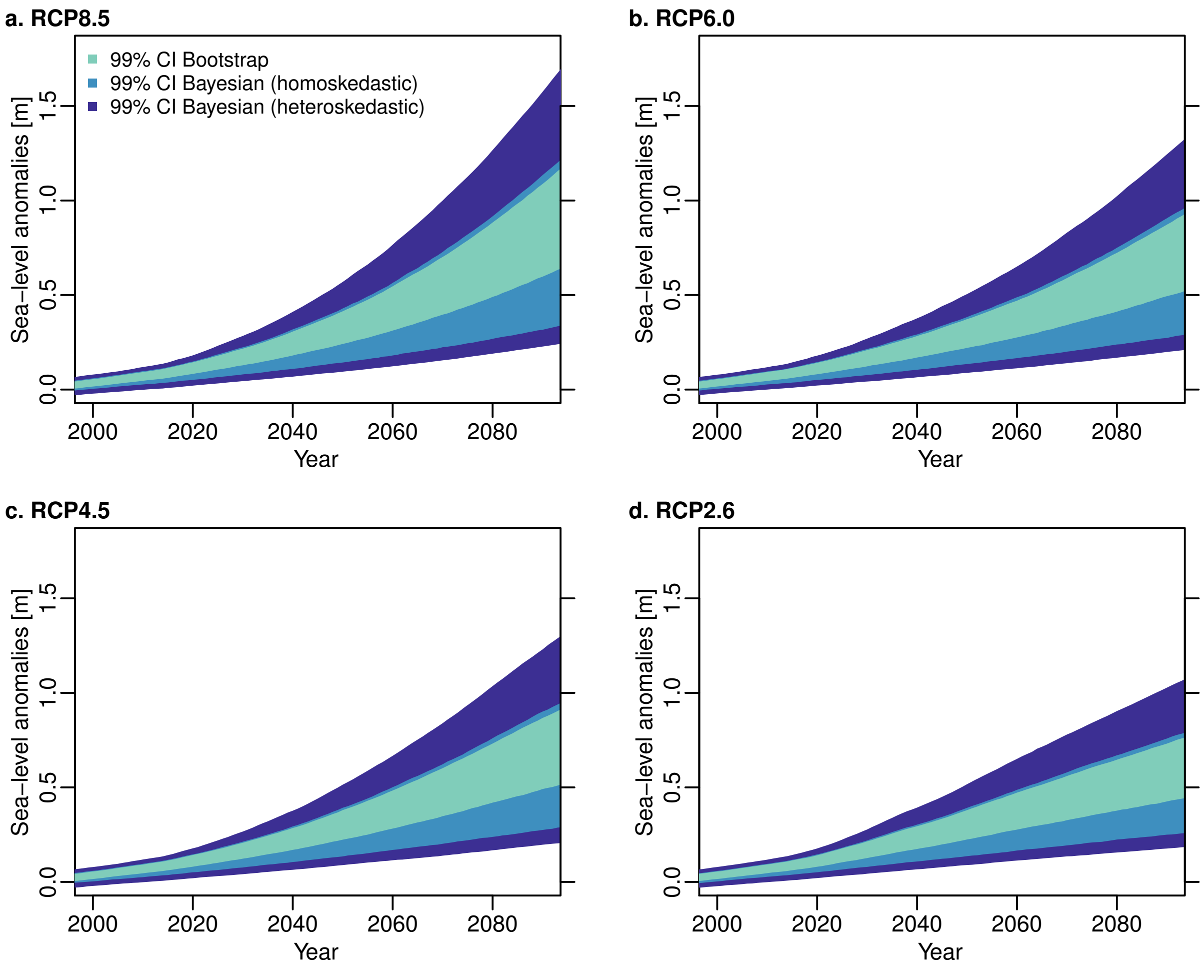

Figure 12: **Comparison of 99% credible intervals of global mean sea-level rise for the three methods using the four RCP temperature scenarios and the Rahmstorf [14] model.** The RCP temperature scenarios are created with the CMIP5, HadGEM2-ES model. The colored envelopes represent plausible estimates from varying methods within the 99% credible interval for (a) RCP8.5, (b) RCP6.0, (c) RCP4.5, and (d) RCP2.6.

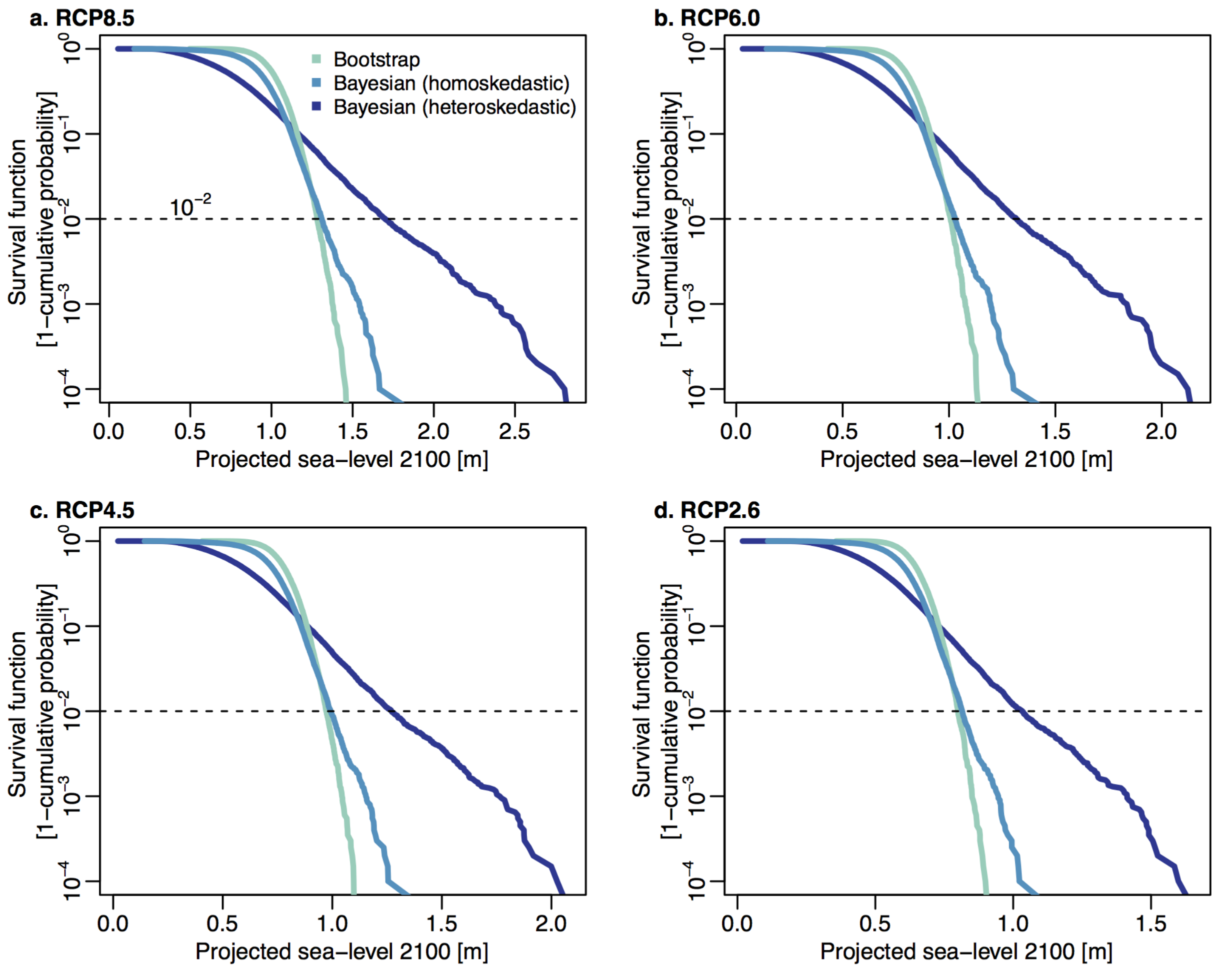

Figure 13: **Comparison of the survival function for global mean sea-level rise in 2100 determined for the different methods using four RCP temperature scenarios and the Rahmstorf [14] model.** The RCP temperature scenarios are created with the CMIP5, HadGEM2-ES model ((a) RCP8.5, (b) RCP6.0, (c) RCP4.5, and (d) RCP2.6). The horizontal dashed lines represent the 1% (10^{-2}) probability in the year 2100.