IMPROVING THE STATISTICAL METHOD
CAN RAISE THE UPPER TAIL OF
SEA-LEVEL PROJECTIONS

A Thesis in
Geosciences

by
Kelsey L. Ruckert

© 2015 Kelsey L. Ruckert

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

December 2015
The thesis of Kelsey L. Ruckert was reviewed and approved* by the following:

Klaus Keller  
Professor of Geosciences  
Thesis Advisor

Chris E. Forest  
Associate Professor of Climate Dynamics

James F. Kasting  
Evan Pugh Professor of Geosciences

Michael Arthur  
Professor of Geosciences  
Interim Associate Head for Graduate Programs in Geosciences

*Signatures are on file in the Graduate School.
Abstract

Anthropogenic greenhouse gas emissions have caused increases in temperatures and sea-levels. Currently, many large urban cities reside within meters of present-day sea-level (e.g., Miami, New York City, New Orleans, Tokyo, Amsterdam, and Mumbai). To protect people and infrastructure against catastrophic events, many cities implement strategies to adapt to sea-level rise. These strategies are often designed for low annual flooding probabilities (i.e., one in 50 to one in 10,000). As a result, the design of these strategies hinges on the upper tails of sea-level rise projections.

Over the past decades, studies have used various calibration methods including frequentist bootstrap and Bayesian inversion to provide probabilistic projections. However, these studies often do not estimate explicitly the upper tail and neglect the effect of known properties of observations. Sea-level observations often have autocorrelated (interdependent) residuals (model minus data) and typically have time-dependent (heteroskedastic) observation errors. In this thesis, we analyze the impact of neglecting such known observed properties on sea-level projections. Specifically, we compare the output of a semi-empirical sea-level model calibrated with different statistical estimation methods: (i) a frequentist bootstrap, (ii) a Bayesian inversion neglecting heteroskedastic residuals, and (iii) a Bayesian inversion considering heteroskedastic residuals. We show how implementing a more appropriate calibration method (in terms of accounting for known observed properties of the residuals) leads to a sizeable increase in the upper tail of sea-level rise projections. The results indicate that the choice of the calibration method can have considerable implications for the design of sea-level rise adaptation strategies.
Table of Contents

List of Figures .................................................................................................................... vi
List of Tables ..................................................................................................................... x
Preface ................................................................................................................................ xi
Acknowledgements ........................................................................................................... xii

Chapter 1
Introduction ............................................................................................................................ 1
1.1 Introduction ................................................................................................................... 1
1.2 Why does sea-level rise matter? ................................................................................... 2
1.3 How have previous studies projected sea-level? ......................................................... 3
1.4 Current challenges and ways forward ......................................................................... 5
References ......................................................................................................................... 11

Chapter 2
Improving the statistical method can raise the upper tail of sea-level projections ........... 15
2.1 Abstract ....................................................................................................................... 15
2.2 Introduction ................................................................................................................ 15
   2.2.1 Sea-level model ................................................................................................. 17
   2.2.2 Analyzed calibration methods ......................................................................... 17
2.3 Method implementation .............................................................................................. 18
2.4 Results and discussion .............................................................................................. 20
   2.4.1 Hindcasts and 90% credible intervals ............................................................... 20
   2.4.2 Do the calibration methods produce reasonable surprise indices? ................. 20
   2.4.3 Parameter and projection distributions ............................................................ 21
   2.4.4 Comparison of reliable sea-level estimates ....................................................... 21
2.5 Conclusion ................................................................................................................ 22
References ......................................................................................................................... 27

Chapter 3
Potential Future Work ........................................................................................................ 32
References ......................................................................................................................... 34
Appendix A
Calibration Methods in Detail ................................................................. 36
A.1 Frequentist bootstrap (Method A) ......................................................... 36
A.2 Bayesian inversion using Markov Chain Monte Carlo (Method B & C) .......... 36
A.3 Testing for bias in the surprise index .................................................... 39
References ................................................................................................. 41

Appendix B
Supplementary Material ........................................................................... 42
List of Figures

Figure 1.1: Industrial carbon dioxide emissions from 1750 to 2010. The points correspond to CO₂ emission estimates expressed in million metric tons of carbon as reported by Boden et al. [2010].

Figure 1.2: Historical industrial CO₂ emissions and CO₂ concentrations. Points correspond to CO₂ emission estimates as reported by Boden et al. [2010]. Maroon squares correspond to the Mauna Loa CO₂ record (in ppm) (P. Tans, NOAA/ESRL (www.esrl.noaa.gov/gmd/ccgg/trends/) and R. Keeling, Scripps Institution of Oceanography (scrippsco2.ucsd.edu/)). Blue circles are reconstructions of CO₂ concentrations from the Law Dome ice core [Etheridge et al., 1998].

Figure 1.3: CO₂ concentrations and combined global land and ocean surface air temperatures. Orange triangles are temperatures as reported in Brohan et al. [2005]. Blue circles are reconstructions of CO₂ concentrations [Etheridge et al., 1998]. Maroon squares are CO₂ concentrations (P. Tans, NOAA/ESRL (www.esrl.noaa.gov/gmd/ccgg/trends/) and R. Keeling, Scripps Institution of Oceanography (scrippsco2.ucsd.edu/)).

Figure 1.4: Combined global mean land and ocean surface air temperatures and global mean sea-level anomalies. Orange triangles are global temperatures as reported in Brohan et al. [2005]. Blue points are sea-level anomalies (in m) derived from tide gauges in Church and White [2006].

Figure 1.5: Historical global mean sea-level anomalies and projections. Orange triangles are global temperatures as reported in Brohan et al. [2005]. Blue points are global sea-level anomalies derived from tide gauges in Church and White [2006]. Projections (purple, pink, yellow, and green lines) are generated from the semi-empirical model in Parris et al. [2012]. The purple projection represents the highest potential sea-level scenario. Pink represents the intermediate-high scenario. The yellow projection is the intermediate-low scenario and the green projection represents the lowest sea-level scenario.

Figure 1.6: Sea-level rise at the Battery in New York City. The black line represents the monthly mean sea-level while the blue line represents its linear trend [NOAA, 2013]. The sea-level anomaly is with respect to the 1990 mean sea-level. The orange dashed line represents the sea-level anomaly in 1983 and the red dashed line is the mean sea-level in the year 2013.
Figure 1.7: Comparison of a subset of previous sea-level rise projections. Red blocks show that the study does not include or neglects to specify such assumptions or results in the study. Green blocks indicate the study reports using the assumption or result.

Figure 2.1: Comparison of hindcasts (a) and projections (b) of global mean sea-level rise (SLR) for the three analyzed calibration methods. Shown are the 90 percent confidence interval for (a) hindcasts from 1880 to 2001 and (b) projections. The colored lines show the 90% credible intervals from the considered calibration methods. The points show the synthesized observations [Church and White, 2006]. The dashed lines represent the SLR range reported by Rahmstorf [2007].

Figure 2.2: Reliability diagrams (panel a and b) and surprise index (panel c) for the three calibration methods. The dashed line represents a model that is neither over- nor underconfident. Panel (a) shows the credible interval from 10 to 100% in increments of 10. Panel (b) zooms in on the confidence intervals from 90 to 100%. Above 96% credibility all methods are at 100% coverage and hence underconfident. Panel (c) shows the surprise index or the deviation from a perfect model.

Figure 2.3: Marginal probability density functions of the estimated model parameters. Shown are sensitivity of sea-level to changes in temperature (parameter a, panel a), temperature when the sea-level anomaly is zero (parameter T0, panel b), and initial sea-level anomaly in the year 1880 (parameter H0, panel c). The vertical line represents the best estimate in Rahmstorf [2007] and the dashed lines are the prior parameter distribution used in the Bayesian inversion (Methods B and C).

Figure 2.4: Projections for global mean SLR in 2050 determined for the four different methods presented as the probability density function (panel a), the cumulative density function (panel b), and the survival function (panel c). The horizontal dashed lines in the survival function (panel c) represent the reliability of 1:100 and 1:10,000 used in example studies to design SLR adaptation strategies [IWR, 2011; Houston, 2013].

Figure B1: Best hindcasts of global mean sea-level observations for the four considered methods. The median fit for the calibration 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 in Church and White [2006]. The points show the
synthesized observations in relation to the 1990 global mean sea level [Church and White, 2006].

Figure B2: Demonstration of the heteroskedastic properties of the observations (panel a) and the autocorrelations of the data-model residuals (panel b) in the global mean sea level record. Panel a shows the errors are time-dependent in the synthesized global mean sea level observations from Church and White [2006]. In panel b the autocorrelation coefficient for the residuals (observations – model simulation). The dashed blue lines indicate the 95% significance of autocorrelation. The vertical lines at the time lags exceed the dashed blue lines meaning the residuals are autocorrelated.

Figure B3: Marginal probability density functions of the estimated model parameters for Method A using five different random number seeds. The overlays of the different realizations are virtually indistinguishable indicating that the chain has converged. Seed 1 represents the seed used in this analysis and the dashed lines are the prior parameter distribution used in the bayesian inversion (Methods B and C).

Figure B4: Marginal probability density functions of the estimated model parameters for Method B using five different random number seeds. The overlays of the different realizations are virtually indistinguishable indicating that the Markov chain has converged. Seed 1 represents the seed used in this analysis and the dashed lines are the prior parameter distribution used in the bayesian inversion (Methods B and C).

Figure B5: Marginal probability density functions of the estimated model parameters for Method C using five different random number seeds. The overlays of the different realizations are virtually indistinguishable indicating that the Markov chain has converged. Seed 1 represents the seed used in this analysis and the dashed lines are the prior parameter distribution used in the bayesian inversion (Methods B and C).

Figure B6: Comparison of 90% (panel a) and 99% (panel b) credible intervals of global mean SLR for the three calibration methods. The colored lines represent plausible estimates from varying calibration methods within the 90% credible interval (panel a) and the plausible estimates within the 99% credible interval (panel b). The points show the synthesized observations of SLR historical measurements estimated from tide gauges. The dashed lines illustrate the minimum and maximum projection of SLR as reported in Rahmstorf [2007].
Figure B7: Comparison of three 90% credible intervals generated in a perfect model experiment by varying assumptions about the observation properties. The gold envelope represents the 90% credible interval generated by assuming 200 independent and identical distributed (i.i.d.) residuals. The gray envelope is the 90% credible interval generated from an AR(1) process and accounts homoskedastic errors. The blue envelope denotes the 90% credible interval produced by accounting for heteroskedastic errors and the AR(1) process. The points are the synthesized observations and the red lines represent the generated observational errors.

Figure B8: Test of the impact of increasing the number of observations on the probability density function (parameter a, panel a & parameter b, panel b) and reliability diagram (panel c & d) in the simple linear test case. 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 (1-200) used in calibration.

Figure B9: Test of the impact of varying assumptions about the observation properties on the probability density function (parameter a, panel a & parameter b, panel b) and reliability diagram (panel c & d) in the simple linear test case. 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 lines and points represent calibration with different assumptions about the observation properties; gold as 200 i.i.d. residuals, gray as autocorrelated residuals, and blue as autocorrelated and heteroskedastic residuals.

Figure B10: Projections for global mean SLR in 2100 determined for the four methods presented as a probability density function (panel a), a cumulative density function (panel b), and a survival function (panel c). The horizontal dashed lines in the survival function (panel c) represent the reliability of 1:100 and 1:10,000 used in example studies to design SLR adaptation strategies [IWR, 2011; Houston, 2013].
List of Tables

Table 1.1: Assumptions of the three analyzed calibration methods. A check mark means the calibration method considers the assumption. ........................................................................................................... 10
Table B1: Prior distributions and calibrated minimum, median, and maximum estimates for model parameters ........................................................................................................................................... 52
Preface

This thesis includes a manuscript projected for submission to a peer-reviewed journal. Kelsey L. Ruckert, the candidate for Master of Science, is the first author. The thesis advisor, Klaus Keller, is the last author, initiated the study, and is the corresponding author. Kelsey L. Ruckert performed the analysis and drafted the paper. Yawen Guan provided and wrote the MCMC likelihood codes. Chris E. Forest and Klaus Keller provided logistical support, general oversight of this study, insightful comments, and stimulating discussions. All authors contributed to the study design, discussed the results, and edited the manuscript.
Acknowledgements

We thank Stefan Rahmstorf for providing his global sea-level model, (http://www.sciencemag.org/content/317/5846/1866.4/suppl/DC1), the scientists who gathered and complied the observations used in this study, Alexander Bakker for testing the reproducibility of the code, and Patrick Applegate for comments on draft versions of this paper.

This work was partially supported by the U.S. Department of Energy, Office of Science, Biological and Environmental Research Program, Integrated Assessment Program, Grant No. DE-SC0005171 with additional support from the National Science Foundation through the Network for Sustainable Climate Risk Management (SCRiM) under NSF cooperative agreement GEO-1240507, and the Penn State Center for Climate Risk Management. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Department of Energy.

Special thanks goes to the entire SCRiM team and Forest Research group for thought-provoking discussions and essential insights during the weekly meetings. I thank Patrick Applegate and Gregory Garner for their advice with the R programming language.

I am greatly indebted to my advisor, Klaus Keller, for his patience, encouragement, and advice. He has taught me many valuable lessons (professionally, personally, and about German culture) and I am deeply grateful for all he has done for me. I thank my committee members for the guidance of the study.

Finally, I am very thankful to my family members and friends for their support and encouragement, especially to my sister who provided me with housing. Mom, Dad, Alison, Bobby, and Rob…thank you.
1.1 Introduction

Humans are emitting greenhouse gases and subsequently influencing sea-level rise. Human emissions of carbon dioxide come from activities such as cement production, deforestation, and burning of fossil fuels (coal, oil, and natural gas) [Hansen et al., 1981]. Since the Industrial Revolution (about 1760-1840), humans have been increasing their emission of carbon dioxide (Fig. 1.1) [Boden et al., 2009]. As a result of increasing carbon dioxide emissions, the amount of carbon dioxide in the atmosphere has increased (Fig. 1.2) [Keeling et al., 1976; Thoning et al., 1989; Etheridge et al., 1998].

The changing concentration of the greenhouse gas carbon dioxide has affected the global climate through the greenhouse effect [Fourier, 1827; Arrhenius, 1896]. Radiation from the sun warms Earth’s surface. The Earth absorbs some of this energy and emits some back to space [Gill, 1982; Barkstrom and Smith, 1986]. In the atmosphere, greenhouse gases absorb the infrared radiation leaving Earth’s surface causing further warming of the planet [Hansen et al., 1981]. As society increased the carbon dioxide concentration in the atmosphere, they increased the absorption of infrared radiation in the atmosphere [Hansen et al., 1981]. This increased rate of absorption led to an increase in the global temperature above the land and ocean (Fig. 1.3) [Brohan et al., 2005; Smith et al., 2008].

The warming temperatures above the land and ocean impact sea-levels. Changes in sea-level are driven by several factors including (i) changes in the density of seawater, (ii) thermal expansion, (iii) land movements (i.e., subsidence, uplift, and tectonics), (iv) changes in terrestrial water storage, and (v) variations in the size of glaciers and ice sheets [Church et al., 2013]. Global and regional temperatures primarily impact the size of glaciers and ice sheets in addition
to thermal expansion. Over the past two centuries, tide gauges around the world show that sea-levels are generally rising [Church and White, 2006; Jevrejeva et al., 2008]. In particular, global mean sea-level rose by roughly 25 cm over the last two centuries [Jevrejeva et al., 2008]. The rise in global mean sea-level coincides with the rise in carbon dioxide and global temperatures (Fig. 1.4) [Solomon et al., 2009]. Current projections suggest that carbon dioxide and temperature will continue to rise with human emissions [Solomon et al., 2009; IPCC, 2013]. If temperatures continue to rise, the impact could potentially cause rapid melting of the ice sheets and glaciers [Applegate et al., 2012; Church et al., 2013]. Several studies project further increase of sea-level rise and at a potentially accelerating rate (Fig. 1.5) [Rahmstorf, 2007; Grinsted et al., 2009; Parris et al., 2012; Kopp et al., 2014].

1.2 Why does sea-level rise matter?

Sea-level rise is a concern for society since it increases the threat for flooding. Today, two-thirds of the world’s largest cities, 13% of all urban population, and approximately 10% of the world’s population (roughly 700 million people) reside within 10 m of sea-level [McGranahan et al., 2007]. The growing flooding threat has motivated analyses of risk management strategies. However, protective measures for sea-level adaptation are costly and take time to implement [Rosenzweig and Solecki, 2014]. Designing these strategies hinges on the resolution of flooding probabilities of both sea-level and storm surges, typically for low values such as from one in 50 to one in 10,000 [IWR, 2011; Houston, 2013]. Quantifying the probability of global and regional annual sea-level can inform the design of risk management strategies.

Consider the example about the importance of quantifying regional annual sea-level in New York City. On the evening of October 29, 2012, Hurricane Sandy struck New York City. Along the coastal regions of New York City, the landfall of Hurricane Sandy coincided with high tide generating a peak storm surge of 2.9 m [Rosenzweig and Solecki, 2014]. This storm surge was large enough to cause a system-wide shutdown of the Metropolitan Transit Authority and the Department of Transportation, inundate about 16% of Staten Island (affecting more than 75,000 people), cause 43 deaths, and generate $19 billion USD of damages in New York City [Rosenzweig and Solecki, 2014]. One reason for the devastation of Hurricane Sandy is because
the storm surge was on top of an increase in sea-level over the last century (Fig. 1.6) [NOAA, 2013].

When Hurricane Sandy made landfall in New York City, risk management strategies had been adapted to the 100-year and 500-year floodplain in the 1983 flood insurance rate maps (FIRMs) [FEMA, 2013; SIRR, 2013]. In the year 2010, the 100-year floodplain from the 1983 FIRMs contained an area of 85 km², which put all 14 of the city’s waste water treatment plants, 12 power plants, 35,500 commercial buildings, and 218,000 people at risk [SIRR, 2013].

Since 1983, there has been considerable development along the city’s shoreline in addition to an increase in sea-levels [NOAA, 2013; SIRR, 2013]. At the Battery, sea-level increased roughly 10 cm (0.1 m) from 1983 to the year 2013 (when flood frequencies were reassessed) (Fig. 1.6) [NOAA, 2013]. The reassessed 100-year floodplain contains 124 km² and 400,000 people at risk [SIRR, 2013]. In the future, sea-level at the Battery is projected to increase by 28 and 79 cm (90th percentile) in the year 2020 and 2050, respectively [SIRR, 2013]. This projected sea-level rise renders current flood frequency maps outdated. As sea-levels rise, the return period decreases and the flooding area increases allowing for storm surges to flood further inland and more frequently [SIRR, 2013]. Hurricane Sandy demonstrates the importance of the resolution of low flooding probabilities and regularly updating coastal flood frequency maps [SIRR, 2013]. Additionally, the damages caused in the wake of Hurricane Sandy motivates the issue of urban resilience to climate risks, especially sea-level rise, not only in New York City, but in cities around the world [Rosenzweig and Solecki, 2014].

1.3 How have previous studies projected sea-level?

Many studies within the last decade focus on projecting global and regional sea-level rise. These studies provide important insights to risk management strategies and on future changes. Over time, these studies have built upon one another to provide more appropriate projections (Table B1).

At various locations around the world, tide gauges are measuring rises in sea-levels. However, estimating rises and acceleration in global mean sea-level is nontrivial [see, for
example Church et al., 2004; Jevrejeva et al., 2006, 2008; Church and White, 2006, 2011]. The studies concluded that global mean sea-level acceleration appears to have started at the end of the 18th century and project sea-level to continue rising at an accelerating rate [Church et al., 2004; Jevrejeva et al., 2006, 2008; Church and White, 2006, 2011].

The next challenge is to project future changes in global mean sea level. One important paper projecting sea-level rise is Rahmstorf [2007]. In Rahmstorf [2007], a sea-level model was proposed by approximating sea-level as a linear response to changes in temperature. In projecting the model with multiple potential future temperature scenarios, the study concluded global mean sea-level could potentially increase over 1 m from 1990 to the year 2100 [Rahmstorf, 2007].

Expanding on the work in Rahmstorf [2007], Grinsted et al. [2009] and Rahmstorf et al. [2012] use a similar model to project sea-level. However, these studies expand on projecting sea-level rise by using an improved model structure and/or statistical calibration method. For example, Grinsted et al. [2009] accounts for the fact that model residuals (model minus data) are interdependent (autocorrelated) and constructs probabilistic projections. Rahmstorf et al. [2012] accounts for the non-constant (heteroskedastic) observation errors.

A more recent study [Kopp et al., 2014] provides local resolved sea-level rise projections. This is an important step to be more relevant to adaptation decisions. Designing local adaptation strategies also hinges on projections on a timescale relevant for infrastructure design. In order to represent local sea-level, Kopp et al. [2014] project the key drivers of sea-level rise: changes in ice sheets (the Greenland ice sheet, the West Antarctic ice sheet, and the East Antarctic ice sheet), glacier and ice cap surface mass balance, global mean thermal expansion and regional ocean effects, land water storage, glacial isostatic adjustment, sediment compaction, and tectonics [Kopp et al., 2014]. This study applies an advanced data-model fusion approach to provide probabilistic projections of each component and global sea-level rise. Additionally, the study provides tail estimates (the 1:100 level) for several local sites of interest.
1.4 Current challenges and ways forward

Despite the recent progress in our ability to project sea-level, deep uncertainties still remain. A wide range of methods and assumptions are used for the model calibration step (Fig. 1.7). The concern in designing adaptation strategies is that studies often choose simplified calibration methods or neglect known uncertainties that can potentially lead to a decrease in the upper tail of sea-level rise projections [e.g., Sriver et al., 2012]. The resolution of the upper tail is critical to the design of many coastal flood protections [NOHSEP, 2011; SIRR, 2013; Van Alphen, 2014]. The current study aims to quantify how a better representation of known observational data properties (heteroskedastic and autocorrelated residuals) can improve probabilistic hindcasts (in terms of the surprise index) and affect the upper tail of sea-level rise projections.

This thesis analyzes the impact of methodological choice about how to fuse observational sea-level changes with a simple model of sea-level projections. In particular, we calibrate a semi-empirical sea-level model with three statistical estimation methods. Each method differs by their representation of observed properties (Table 1.1). The first method, a frequentist bootstrap neglects the effect of heteroskedastic residuals and assumes a known autocorrelation coefficient. The second method, a Bayesian inversion still neglects the heteroskedastic properties, but accounts for uncertainty in the autocorrelation process. The last considered method is a Bayesian inversion that accounts for both heteroskedastic properties and uncertainty in the autocorrelation process. In comparing the results, we address three questions: (i) Does the choice of calibration method impact projections?, (ii) Does neglecting observed properties (heteroskedastic errors and uncertainty in the autocorrelation process) cause a decrease in the upper tail estimates?, and (iii) Which of the considered methods is the best to calibrate model parameters?
**Figure 1.1:** Industrial carbon dioxide emissions from 1750 to 2010. The points correspond to CO₂ emission estimates expressed in million metric tons of carbon as reported by Boden et al. [2010].

**Figure 1.2:** Historical industrial CO₂ emissions and CO₂ concentrations. Points correspond to CO₂ emission estimates as reported by Boden et al. [2010]. Maroon squares correspond to the Mauna Loa CO₂ record (in ppm) (P. Tans, NOAA/ESRL (www.esrl.noaa.gov/gmd/ccgg/trends/) and R. Keeling, Scripps Institution of Oceanography (scrippsco2.ucsd.edu/)). Blue circles are reconstructions of CO₂ concentrations from the Law Dome ice core [Etheridge et al., 1998].
Figure 1.3: CO₂ concentrations and combined global land and ocean surface air temperatures. Orange triangles are temperatures as reported in Brohan et al. [2005]. Blue circles are reconstructions of CO₂ concentrations [Etheridge et al., 1998]. Maroon squares are CO₂ concentrations (P. Tans, NOAA/ESRL (www.esrl.noaa.gov/gmd/ccgg/trends/) and R. Keeling, Scripps Institution of Oceanography (scrippsc02.ucsd.edu/)).

Figure 1.4: Combined global mean land and ocean surface air temperatures and global mean sea-level anomalies. Orange triangles are global temperatures as reported in Brohan et al. [2005]. Blue points are sea-level anomalies (in m) derived from tide gauges in Church and White [2006].
Figure 1.5: Historical global mean sea-level anomalies and projections. Orange triangles are global temperatures as reported in Brohan et al. [2005]. Blue points are global sea-level anomalies derived from tide gauges in Church and White [2006]. Projections (purple, pink, yellow, and green lines) are generated from the semi-empirical model in Parris et al. [2012]. The purple projection represents the highest potential sea-level scenario. Pink represents the intermediate-high scenario. The yellow projection is the intermediate-low scenario and the green projection represents the lowest sea-level scenario.
Figure 1.6: Sea-level rise at the Battery in New York City. The black line represents the monthly mean sea-level while the blue line represents its linear trend [NOAA, 2013]. The sea-level anomaly is with respect to the 1990 mean sea-level. The orange dashed line represents the sea-level anomaly in 1983 and the red dashed line is the mean sea-level in the year 2013.
**Figure 1.7:** Comparison of a subset of previous sea-level rise projections. Red blocks show that the study does not include or neglects to specify such assumptions or results in the study. Green blocks indicate the study reports using the assumption or result.

**Table 1.1:** Assumptions of the three analyzed calibration methods. A check mark means the calibration method considers the assumption.
References


IWR (2011), Flood Risk Management Approaches: As Being Practiced in Japan, Netherlands, United Kingdom, and United States, United States Army Corps of Engineers Institute for Water Resources (IWR). IWR-2011-R-08.


1 Improving the statistical method can raise the upper tail of sea-level projections

2.1 Abstract

Strategies to manage risks driven by sea-level rise (SLR) are often designed for low annual flooding probabilities such as 1:50 to 1:10,000. This high reliability requires information about the upper tails of SLR projections. Previous SLR projections have provided important insights. However, they often do not report the upper tail and are silent on the effects of time-varying (heteroskedastic) observation errors and uncertain autocorrelated residuals. Here we use a previously published semi-empirical sea-level model to quantify the impacts of these effects. We compare three calibration methods: a frequentist bootstrap as well as a Bayesian inversion neglecting and considering heteroskedastic residuals. The Bayesian inversion accounting for heteroskedasticity produces a hindcast with reasonable surprise indices and accounts best for observed properties of the residuals. Choosing this more appropriate method increases, in our analysis, the 1:100 and 1:10,000 reliable SLR in 2050 by 0.09 and 0.32 meters compared to a frequentist bootstrap.

2.2 Introduction

Anthropogenic greenhouse gas emissions have caused global warming and SLR during the last century [Church and White, 2006; Grinsted et al., 2009]. Because rising sea-levels drive increases in flood risks [McGranahan et al., 2007; Rosenzweig and Solecki, 2014], future flooding risks hinge critically on sea-level projections [Houston, 2013].

1 This chapter is projected for submission for publication at Environmental Research Letters. For details about the authors please view the preface on page xii.
Sea-level projections are often derived by calibrating semi-empirical sea-level models to inform risk and decision analyses [Meehl et al., 2007, Rahmstorf 2007]. For example, Herberger et al. [2009], McInnes et al. [2013], and Neumann et al. [2011] identify areas vulnerable to coastal flooding and probable cost at risk by adding a ‘high-end’ estimate of SLR and today’s 100-year storm tide height to maps of the California, Australia, and U.S. east coast. These studies use the reported ‘high-end’ SLR estimates for 2100 from the IPCC AR4 and Rahmstorf [2007]. The IPCC AR4 assesses SLR based on various future greenhouse gas emissions and the combination of the contributions from thermal expansion of the oceans, melting of the glaciers and ice sheets, and the potential of rapid dynamic ice-sheet response [Meehl et al., 2007]. Herberger et al. [2009] and Neumann et al. [2011] adjust the IPCC AR4 and Rahmstorf [2007] values to represent site-specific SLR by accounting for water trapped in dams and reservoirs or changes in land height. As a last example, consider the approach taken by the US Army Corps of Engineers (USACE) and the Shoreline Management Plan (SMP). Both references use the ‘high-end’ estimate reported in Rahmstorf [2007] for the design of SLR strategies [Convertino et al., 2010; Dalton et al., 2010; Dibajnia et al., 2012].

These studies that result in real-world risk and decision-analysis often choose simplified calibration methods that potentially neglect autocorrelation, heteroskedastic residuals, and/or low probabilities. Many studies projecting SLR focus on the 90% (5-95%) credible interval [Grinsted et al., 2009; Ezer and Corelett, 2012; Jevrejeva et al., 2012; Rahmstorf et al., 2012; Grassi et al., 2013; Moore et al., 2013; Perrette et al., 2013]. However, decision makers are often interested in tail probabilities far beyond the 90% credible interval. Infrastructures that manage flood risks are often designed to withstand very low annual flood frequencies around 1:50 to 1:10,000 [Jonkman et al., 2009; IWR, 2011]. Designing such infrastructure requires careful resolution of low probabilities from both storm surges and SLR [Jonkman et al., 2009; Houston, 2013].

Probabilistic projections rely heavily on assumptions about the statistical model used for calibration. Key challenges include (i) the representation of non-constant (heteroskedastic) data-model residual errors (Fig. B1 & B2a), (ii) the importance of interdependent (autocorrelated)
data-model residuals (Fig. B2b), and (iii) the representation of low probability/high impact events in the tail [Zellner and Tiao, 1964; Ricciuto et al., 2008]. Past SLR projections frequently assume homoskedastic (constant) residuals [Rahmstorf, 2007; Grassi et al., 2013; Moore et al., 2013; Kopp et al., 2014]. Assuming homoskedastic errors neglects the potential for heteroskedastic observation errors. Neglecting these observation properties can potentially lead to overconfidence [e.g., Zellner and Tiao, 1964; Ricciuto et al., 2008].

Here we show how a better representation of autocorrelation and heteroskedastic residuals can improve the probabilistic hindcasts (in terms of the surprise index) and lead to a sizeable rise in the upper tail of global mean SLR projections.

2.2.1 Sea-level model

To demonstrate the effect, we adopt the widely used semi-empirical model projecting SLR presented in Rahmstorf [2007] (Fig. 2.1). The Rahmstorf [2007] model uses global mean temperatures (T) and an estimate of global mean sea-level derived from tide gauge observations to predict SLR,

\[
\frac{\partial H}{\partial t} = \alpha (T - T_0),
\]

(1)

where \( \alpha \) is sensitivity of sea-level to temperature, \( T \) global mean surface air temperatures, \( T_0 \) the temperature when sea-level is zero, \( H \) the global mean sea-level, and \( t \) is time. We slightly expand on the calibration setup by including the initial value of SLR in 1880 (H_0) as an uncertain parameter. We adopt the same data source from Rahmstorf [2007] for sea-level observations [Church and White, 2006]. Sea-level observations are anomalies with respect to the average 1990 sea-level and temperature anomalies with respect to the 20th century. We update the data sources for historic temperatures [Smith et al., 2008] and use the bias corrected CMIP5, CNRM-CM5 RCP 8.5 temperature scenario for projections to the year 2100 [Voldoire et al., 2011; Taylor et al., 2012].

2.2.2 Analyzed calibration methods

We compare parameter estimates and SLR projections based on three commonly used calibration methods: (A) a bootstrap method [Solow, 1985; Rizzo, 2007], (B) a Bayesian method
assuming homoskedastic errors, and (C) a Bayesian method accounting for the observed time dependent (heteroskedastic) observation error (details in appendix) [Metropolis et al., 1953; Zellner and Tiao, 1964; Hastings, 1970; Gilks, 1997]. Method A (Bootstrap) produces replicates of the residuals using a best-fit, time-series model [Solow, 1985; Rizzo, 2007; Tonidandel et al., 2009; Lempert et al., 2012]. Methods B and C use a Bayesian inversion to estimate the posterior density using a Markov Chain Monte Carlo method and the Metropolis Hastings algorithm [Metropolis et al., 1953; Zellner and Tiao, 1964; Hastings, 1970; Gilks, 1997]. Method B approximates the residuals as normal and independent with zero means and a common variance (homoskedastic errors). Method C accounts for the time-varying observational error, which generally decrease over time in the sea-level record (Fig. B2). All three methods retain the autocorrelated structure of the residuals (Fig. B2). Method A approximates this with the most likely value, while methods B and C account for uncertainty in the autocorrelation parameters. All three calibration methods use the same temperature and sea-level data covering the time period of 1880 to 2002.

2.3 Method implementation

The calibration methods approximate the observation dataset as the sum of the model output plus a residual term:

\[ y_{\text{observations}} = f(\theta, t) + R_t. \]  

The Bayesian inversion in Method B and C uses uniform prior distributions for the model parameters (\( \theta \)) (Table B1). Additionally, the likelihood function for Method B and C incorporates the variance of the autocorrelation process and the type of observation error (homoskedastic or heteroskedastic) (Table B1). In Method B, the observation errors are set to zero to represent the homoskedastic assumption, whereas the errors are set as the reported values in Method C [Zellner and Tiao, 1964]. We use \( 2 \cdot 10^4 \) (Method A), \( 1 \cdot 10^7 \) (Method B), and \( 2.5 \cdot 10^7 \) (Method C) iterations. Several studies suggest an iteration number \( \geq 1,000 \) is appropriate to produce confidence intervals with Method A [Efron and Tibshirani, 1993; Cole, 1999; Sohn and Menke, 2002]. For Method B and C we remove a one percent initial “burn-in” from the Markov chains [Gilks, 1997]. Additionally, we reduce the chain length in Method B and C to subsets of \( 2 \cdot 10^4 \) for plotting hindcasts and projections by saving every 495\(^{\text{th}}\) and 1237\(^{\text{th}}\)
iteration, respectively. To assess convergence, we run each method with five different random number seeds (numbers initializing the generator) to check for a potential lack of convergence. The different realizations are virtually indistinguishable (Fig. B3, B4, & B5). The hindcasts and projections display the 90% confidence interval (Method A) and the 90% credible interval (Method B and C) for comparison to other sea-level studies (for simplicity we refer to these intervals as credible intervals in the remainder of this paper). Since all methods generate random numbers, we run each method with five different random number seeds to assess the robustness of the main conclusions to this source of variability (Fig. B3, B4, & B5).

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 [Tippett et al., 2014]. Simply put, a highly reliable projection is neither over-nor-underconfident. Visually, this is assumed 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:

$$AC = 100\% \cdot \left[ 1 - \left( \frac{\text{# of data outside the credible interval}}{\text{# of total data points}} \right) \right].$$  \tag{3}

Additionally, we evaluate the skills (measure of performance) of the calibration 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 [Scotti, 2013]. 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{4}

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.
2.4 Results and discussion

2.4.1 Hindcasts and 90% credible intervals

The three considered calibration methods produce visually quite similar hindcasts, but different projections (Fig. 2.1a & 2.1b). Each hindcast passes through more than half the data and the range of observation errors (Fig. B1). The projections out to 2050 differ by up to 0.13 m in range between the 90% credible intervals from the considered calibration methods (Fig. 2.1b). The credible intervals increase from Method A, B, to C. Method C yields a wider 90% credible interval than the total range (minimum to maximum SLR projection) reported in Rahmstorf [2007]. The trend in the ranking from tightest to widest projections is the same for the 90 and 99% credible intervals and increases with projection time up to the year 2100 (Fig. B6).

2.4.2 Do the calibration methods produce reasonable surprise indices?

The three calibration methods share a roughly comparable reliability diagram that predicts reasonably well at low to intermediate credible intervals (Fig. 2.2a). However, the methods differ considerably at the higher credible intervals referenced for high reliabilities (Fig. 2.2b). Overall each method produces a small average deviation (2-4%) from the perfect 1:1 line (Fig. 2.2c). Above the 90% credible level all considered methods are underconfident (Fig. 2.2b).

Possible causes for this underconfidence include the small number of observations, autocorrelated residuals, heteroskedastic residuals, a large number of parameters, priors, and structural model errors. We perform a perfect model experiment with a known structure (details in the appendix) (Fig. B7). In the experiment, we vary the sample size of observations and the assumptions about the observation properties (i.i.d. versus autocorrelated residuals and homoskedastic versus heteroskedastic errors). Consistent with the results of previous studies [Richardson, 2001; Tippett et al., 2014], we find that increasing the number of observations tightens the parameter distribution and visually improves the reliability (Fig. B8). However, we also observe a tightening of parameter distribution and improved reliability moving from autocorrelated and heteroskedastic residuals, to i.i.d. residuals, and to an autocorrelated and homoskedastic case (Fig. B7 & B9). Moving from a homoskedastic to a heteroskedastic case
adds one more parameter to the model hence widening parameter distributions and reducing reliability.

2.4.3 Parameter and projection distributions

Choosing different calibration methods causes considerable differences in the estimated modes and tails of the model parameters (Fig. 2.3 & Table B1) and model projections (Fig. 4). Better accounting for the observed data properties (i.e., moving from Method A to B to C) causes wider parameter estimates (Fig. 2.3 & Table B1), projections, and a decrease in the most likely value (mode) (Fig. 2.4 & Fig. B10). All three calibration methods generate a wider distribution in comparison to the 2050 reported range in Rahmstorf [2007] (0.25 – 0.41 m). Depending on the choice of calibration method, the projected reliable SLR in 2050 can range from 0.35 – 0.44 m (1:100 level) and 0.39 – 0.71 m (1:10,000 level) above the 1990 sea-level (Fig. 2.4c).

2.4.4 Comparison of reliable sea-level estimates

The choice of calibration method considerably impacts the upper tail-area estimates of SLR projections. If we account for heteroskedastic residuals and uncertainty about the autocorrelated residuals (Method C) the 1:100 and a 1:10,000 levels in 2050 are 0.44 m and 0.71 m, respectively. These estimates are roughly 1.3 and 1.8 times larger compared to Method A, which neglects the heteroskedastic properties and assumes a known autocorrelation coefficient. In 2100, the 1:10,000 level from Method C (2.35 m) differs from Method A (1.17 m) by 1.18 m. In addition, all three calibration methods project 1:100 and 1:10,000 levels at least 0.05 m larger in 2050 and 2100 than the optimal ‘medium-end’ (SRES A2 temperature scenario [IPCC, 2007]) estimates in Rahmstorf [2007] (0.3 m and 0.8 m) and the ‘high-end’ estimate in the IPCC AR5 [Church et al, 2013] (0.8 m in 2100) (Fig. 2.4c & Fig. B10).

The differences in the SLR projections and parameter distributions can be traced back to assumptions embedded in the calibration methods. Method A neglects uncertainty in the autocorrelation coefficient and neglects the heteroskedastic nature of the observation errors. Method B still neglects the heteroskedastic nature of the observation errors, but accounts for uncertainty about the autocorrelation coefficient. Our simple analysis quantifies how neglecting
uncertainty about the autocorrelation coefficient and/or assuming homoskedastic errors can lead to overconfident projections.

2.5 Conclusion

Semi-empirical SLR models have often been used to project sea-level changes. Many of these studies do not account for the autocorrelated and heteroskedastic properties of the model residuals. Accounting for these properties (and their associated uncertainties) improves the model hindcasts and reliabilities. We show how the methodological improvements increase the projected SLR in 2050 by a factor of 1.3 with a 1:100 level and almost double the projected SLR with a 1:10,000 level.
**Figure 2.1:** Comparison of hindcasts (a) and projections (b) of global mean sea-level rise (SLR) for the three analyzed calibration methods. Shown are the 90 percent confidence interval for (a) hindcasts from 1880 to 2001 and (b) projections. The colored lines show the 90% credible intervals from the considered calibration methods. The points show the synthesized observations [Church and White, 2006]. The dashed lines represent the SLR range reported by Rahmstorf [2007].
Figure 2.2: Reliability diagrams (panel a and b) and surprise index (panel c) for the three calibration methods. The dashed line represents a model that is neither over-nor-underconfident. Panel (a) shows the credible interval from 10 to 100% in increments of 10. Panel (b) zooms in on the confidence intervals from 90 to 100%. Above 96% credibility all methods are at 100% coverage and hence underconfident. Panel (c) shows the surprise index or the deviation from a perfect model.
Figure 2.3: Marginal probability density functions of the estimated model parameters. Shown are sensitivity of sea-level to changes in temperature (parameter a, panel a), temperature when the sea-level anomaly is zero (parameter T0, panel b), and initial sea-level anomaly in the year 1880 (parameter H0, panel c) The vertical line represents the best estimate in Rahmstorf [2007] and the dashed lines are the prior parameter distribution used in the Bayesian inversion (Methods B and C).
Figure 2.4: Projections for global mean SLR in 2050 determined for the four different methods presented as the probability density function (panel a), the cumulative density function (panel b), and the survival function (panel c). The horizontal dashed lines in the survival function (panel c) represent the reliability of 1:100 and 1:10,000 used in example studies to design SLR adaptation strategies [IWR, 2011; Houston, 2013].
References


IWR (2011), Flood Risk Management Approaches: As Being Practiced in Japan, Netherlands, United Kingdom, and United States, United States Army Corps of Engineers Institute for Water Resources (IWR). IWR-2011-R-08.


Potential Future Work

Our results point to several potential avenues of future research. First, we use a simple three-parameter model for sea-level. It might be of interest to analyze the differences using a more complex sea-level model [Vermeer and Rahmstorf, 2009; Nordhaus, 2010; Kemp et al., 2011; Moore et al., 2013].

Second, the impacts of calibration method choice on projections are potentially important beyond the specific application of sea-level rise. The insights can be relevant to fields of ecology, meteorology, epidemiology, or geology. For example, consider the impact of changes in precipitation [Shongwe et al., 2009, 2011] and subsequently future food security [Lobell et al., 2008]. Additionally, consider the relevance to analyzing the extinction risk of species [Shoemaker et al., 2013]. As a last example, consider the significance to the field of medicine through projecting the spread of infectious disease epidemics [Raftery et al., 2010] and the effectiveness of vaccinations [Wallinga et al., 2003].

One avenue of research we are currently pursuing deals with projecting the contribution of the Antarctic ice sheet to sea-level rise. As first warned by John Mercer [Mercer, 1968, 1978], human emissions of greenhouse gases into the atmosphere could trigger a disintegration of the Antarctic marine ice sheet. This is due to a fast feedback called marine ice sheet instability (MISI) [Ivins, 2009]. MISI is a runaway retreat of the part of the ice sheet that is grounded below sea-level due to the relationship between ice flux and the grounding line [Mercer, 1978; Pollard et al., 2015]. Melting of the ice shelves, which act as dams for glaciers flowing into the ocean, could potentially lead to MISI if the bed deepens towards the continent [Mercer, 1978; Pollard et al., 2015]. Despite the importance of the Antarctic ice sheet to future sea-level change, changes in the volume of the ice sheet are currently rather poorly understood [Ivins, 2009].
To help close the knowledge gap, the current research calibrates the *Shaffer* [2014] Antarctic ice sheet volume loss model (DAIS) with a precalibration (Latin hypercube sampling) and a full Bayesian inversion method. The Bayesian inversion method accounts for the heteroskedastic nature of the data, however the data are independent and identically distributed. This model does account for MISI although the model neglects the effects of cliff instability. Cliff instability is a process that a sheer cliff of ice can only be so tall before it collapses under its own weight. Today this process occurs where the ice shelves calve icebergs [*Pollard et al.*, 2015]. Without this process in the model it could potentially take a longer period of warming or higher temperatures to trigger MISI [*Pollard et al.*, 2015]. In this study, we address three questions: (i) How much does a full Bayesian inversion improve compared to a precalibration?, (ii) Do the calibration methods pass a hindcast test?, and (iii) What are the impacts of the missing fast feedbacks (i.e., cliff instability) on hindcasts and projections?
References


Calibration Methods in Detail

The analyzed calibration methods include a Frequentist bootstrap and a Bayesian inversion accounting for either heteroskedastic or homoskedastic errors. By using the same sea-level data and temperature forcings we evaluate each method’s ability to predict extreme events.

A.1 Frequentist bootstrap (Method A)

The Bootstrap method is based on Monte Carlo simulation that provides confidence intervals by resampling from an observed dataset [Solow, 1985]. We adopt the bootstrap method developed by Solow [1985] for correlated observations. If observed data represents $y_t$, model output $[f(\theta, t)]$, unknown model parameters $\theta$, an unknown residual term $R_t$, and time over a hindcast period of 1880-2002 $t$, then sea-level observations are modeled as the sum of the model output plus an unknown residual term:

$$y_t = f(\theta, t) + R_t.$$  \hfill (A1)

The observations, model outputs, and residuals are vectors from 1 to $i$, a length of 122 years (A2).

$$y_t = \begin{bmatrix} y_{t1} \\ y_{ti} \end{bmatrix} , f_t = \begin{bmatrix} f(\theta, t1) \\ \vdots \\ f(\theta, ti) \end{bmatrix} , R_t = \begin{bmatrix} \omega_{t1} + \epsilon_{t1} \\ \vdots \\ \omega_{ti} + \epsilon_{ti} \end{bmatrix}.$$  \hfill (A2)

The residuals are assumed to be a stationary normal AR(1) first-order autoregressive process ($\omega_t$) with an added observation error ($\epsilon_t$),

$$R_t = \omega_t + \epsilon_t.$$  \hfill (A3)

The observation error in Method A are assumed to be homoskedastic, where the errors consist of a normal (N) distribution ($\sim$) with zero means and a common (constant) variance, $\epsilon_t \sim N(0, \sigma^2_{\epsilon})$. The AR(1) process $\omega_t$ is characterized by the lag-1 (annual) autoregression coefficient as $\rho$, the AR(1) innovation variance $\sigma^2_{AR_1}$, and the white noise as $\delta_t$:

$$\omega_t = \rho \omega_{t-1} + \delta_t.$$  \hfill (A4)
with the initial value \((\omega_0)\) of:

\[
\omega_0 \sim N\left(0, \frac{\sigma^2_{AR1}}{1 - \rho^2}\right).
\]

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 \(f_t\) from the observed data:

\[
R_t = y_t - f_t. \tag{A5}
\]

The best model fit is estimated by minimizing the absolute residuals using a global optimization method (Differential Evolution) [Ardia et al., 2011]. Using the residuals, we estimate the lag-1 (annual) autoregression coefficient \(\rho\) with the autocorrelation function in “R”. The residuals are then resampled with replacement \(i\) times (122 years) to form a bootstrap sample. Taking the standard deviation of the bootstrap sample, we estimate the stationary process variance, \(\sigma^2_{\rho}\). In order to avoid accounting for autocorrelation twice, the stationary process variance is transformed into uncorrelated white noise variance \(\sigma^2_{AR1}\) by taking the square root of the stationary variance times one minus the squared autoregression coefficient:

\[
\sigma^2_{AR1} = \sigma^2_{\rho}(1 - \rho^2). \tag{A6}
\]

Using the white noise variance, the bootstrap simulation for a hindcast is created according to:

\[
y_t = f(\theta, t) + \rho \omega_{t-1} + \delta_t. \tag{A7}
\]

We then estimate model parameters from the bootstrap simulation by minimizing the absolute value of the residuals using Differential Evolution [Ardia et al., 2011]. The calibrated 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 (A7). This process is repeated 20,000 times to create distributions of parameter estimates and SLR.

### A.2 Bayesian inversion using Markov Chain Monte Carlo (Method B & C)

We implement a Bayesian inversion technique using a Markov chain Monte Carlo (MCMC) algorithm [Metropolis et al., 1953; Hastings, 1970; Gilks, 1997]. This Bayesian approach uses observed data to update our prior knowledge about model parameters \(\theta\) to obtain a
distribution of successful (\( \pi \)) parameters as the posterior distribution \( \pi(\theta) \) over a hindcast calibration period from 1880-2002. The overall process follows Bayes’s theorem [Bayes, 1764], where the probability (\( \pi \)) of observing the parameters (\( \theta \)) given (\( R_t \)) the residuals (\( R_t \)) is proportional (\( \propto \)) to the likelihood (\( L \)) of the residuals given the parameters times the prior probability distribution of the parameters,

\[
\pi(\theta | R_t) \propto L(R_t | \theta) \cdot \pi(\theta).
\]

(A8)

Once the prior distribution and likelihood function is defined, the MCMC algorithm samples from the posterior distribution [Metropolis et al., 1953; Hastings, 1970; Gilks, 1997].

We construct the likelihood function, assuming that the observational dataset (\( y_t \)) consists of the model output \( f(\theta,t) \) plus an unknown residual term \( R_t \) (A1 & A2). The residual term also consists of a stationary normal AR(1) first-order autoregressive process (\( \omega_t \)) and observation error (\( \epsilon_t \)) (A3 & A4). 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 \( \theta \) (Supplementary Table 1) and the unknown statistical parameters \( \sigma^2_{AR1} \) (white noise variance) and \( \rho \) (lag-1 autoregressive coefficient). By using an AR(1) likelihood, the method accounts for uncertainty in the autocorrelated process parameters and for autocorrelation in the residuals similar to previous work [Urban and Keller, 2010; Olson et al., 2012]. 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 residuals given the parameters, \( L(R_t | \Theta) \), we find the residuals as in equation A5 and assume they follow a normal distribution, \( R_t \sim N(0, \Sigma) \), where \( \Sigma \) is the sum of the variance of AR(1) process (\( \omega_t \)) and the observation errors (\( \epsilon_t \));

\[
Var(\omega_t) = \frac{\sigma^2_{AR1}}{1-\rho^2} \begin{pmatrix}
1 & \cdots & \rho^{N-1} \\
\vdots & \ddots & \vdots \\
\rho^{N-1} & \cdots & 1
\end{pmatrix},
\]

\[
Var(\epsilon_t) = \begin{pmatrix}
\sigma^2_t & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & \sigma^2_n
\end{pmatrix}
\]

(A9)
\[ \mathbf{\Sigma} = \text{Var}(\omega_t) + \text{Var}(\varepsilon_t) = \frac{\sigma_{AR1}^2}{1 - \rho^2} \begin{pmatrix} 1 & \cdots & \rho^{N-1} \\
 & \ddots & \vdots \\
 & & 1 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & \cdots & 0 \\
 & \ddots & \vdots \\
 & & \sigma_N^2 \end{pmatrix}. \]  
(A10)

The likelihood function is then:

\[ L(\mathbf{R}|\Theta) = \left( \frac{1}{\sqrt{2\pi}} \right)^N |\mathbf{\Sigma}|^{-1/2} e^{-\frac{1}{2} \mathbf{R}^T \mathbf{R} \mathbf{R}^T}. \]  
\[ \Theta = (\theta, \rho, \sigma_{AR1}) \]  
(A11)

We represent two different assumptions about observational errors: homoskedastic (Method B) or heteroskedastic (Method C). In Method B, the observation errors are assumed homoskedastic, where the errors consist of a normal distribution with zero means and a common variance, \( \varepsilon_t \sim N(0, \sigma_\varepsilon^2) \). This means the variance of the observation errors (\( \sigma_\varepsilon^2 \)) is constant through time. In Method C, the observational errors are assumed heteroskedastic, where the errors consist of a normal distribution with zero means and a time dependent variance (\( \sigma_{\varepsilon,t}^2 \)), \( \varepsilon_t \sim N(0, \sigma_{\varepsilon,t}^2) \). The time dependent variance is the measurement errors of the observations provided in Church and White [2006].

We adopt uniform priors for the parameters (Supplementary Table 1). The MCMC uses Metropolis-Hastings updates with joint normal proposal [Metropolis et al., 1953; Hastings, 1970]. We use \( 1 \cdot 10^7 \) and \( 2.5 \cdot 10^7 \) iterations for analysis. The trace of the MCMC chain suggests that the Markov chain is well mixed and converged. We drive the calibrated model with RCP8.5 temperatures to project smooth simulations (\( f_t \)) of SLR [Voldoire et al., 2011; Taylor et al., 2012]. Following equation A7, the smooth simulations are superimposed with the process noise to account for uncertainty in the autocorrelation and the observational errors. The simulations with process noise represent a probabilistic distribution of SLR and the Markov chains represent the distribution of parameter estimates.

A.3 Testing for bias in the surprise index

All three calibration methods produce underconfident hindcasts above the 90% credible interval. This section evaluates the effect of possible explanations for this bias as: (1) number of
observations, (2) autocorrelated observations, or (3) assuming heteroskedastic errors. To analyze this, we perform a perfect model experiment with a known structure for a simple model:

\[ y = a + bx + R_x, \]  

where \( a \) and \( b \) are model parameters, \( x \) time, \( R \) residuals, and \( y \) denotes observations. The residuals consist of either i.i.d. or AR(1) residuals, generated at random. The first test varies the number of observations from 1 to 200 (Supplementary Fig. 5). The second test examines the impact of i.i.d. residuals versus AR(1) residuals assuming homoscedastic errors and AR(1) residuals assuming heteroskedastic errors (Supplementary Fig. 6). The study compares the results based on the parameter distributions and the reliability diagram.

We analyze the impact of varying assumptions about the observation properties on calibration of a sea-level model. The choice of calibration method impacts the parameter distributions and projections. More importantly, this assessment demonstrates how using a more appropriate calibration method, which considers observation properties, increases the projected sea-level rise associated with high reliabilities used to manage flood risks.
References


Appendix B

Supplementary Material

This supporting information provides the 90 and 99% credible interval of sea-level projections for each method to the year 2100. In addition, the supporting information provides (i) a comparison of the median estimate from each method, (ii) an assessment of the time-dependent error in the observations and autocorrelated residuals, (iii) marginal probability density functions of the estimated model parameters for each method using five different random number seeds to assess convergence, (iv) a test of the impacts of changes in the calibration properties for a perfect model experiment, and (v) a table describing the minimum, median, and maximum parameter estimates from each method analyzed in this study.

Figure B1: Best hindcasts of global mean sea-level observations for the four considered methods. The best fit for the calibration 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 in Church and White [2006]. The points show the synthesized observations in relation to the 1990 global mean sea level [Church and White, 2006].
Figure B2: Demonstration of the heteroskedastic properties of the observations (panel a) and the autocorrelations of the data-model residuals (panel b) in the global mean sea level record. Panel a shows the errors are time-dependent in the synthesized global mean sea level observations from Church and White [2006]. In panel b the autocorrelation coefficient for the residuals (observations – model simulation). The dashed blue lines indicate the 95% significance of autocorrelation. The vertical lines at the time lags exceed the dashed blue lines meaning the residuals are autocorrelated.
Figure B3: Marginal probability density functions of the estimated model parameters for Method A using five different random number seeds. The overlays of the different realizations are virtually indistinguishable indicating that the chain has converged. Seed 1 represents the seed used in this analysis and the dashed lines are the prior parameter distribution used in the Bayesian inversion (Methods B and C).
Figure B4: Marginal probability density functions of the estimated model parameters for Method B using five different random number seeds. The overlays of the different realizations are virtually indistinguishable indicating that the Markov chain has converged. Seed 1 represents the seed used in this analysis and the dashed lines are the prior parameter distribution used in the Bayesian inversion (Methods B and C).
Figure B5: Marginal probability density functions of the estimated model parameters for Method C using five different random number seeds. The overlays of the different realizations are virtually indistinguishable indicating that the Markov chain has converged. Seed 1 represents the seed used in this analysis and the dashed lines are the prior parameter distribution used in the Bayesian inversion (Methods B and C).
Figure B6: Comparison of 90% (panel a) and 99% (panel b) credible intervals of global mean SLR for the three calibration methods. The colored lines represent plausible estimates from varying calibration methods within the 90% credible interval (panel a) and the plausible estimates within the 99% credible interval (panel b). The points show the synthesized observations of SLR historical measurements estimated from tide gauges. The dashed lines illustrate the minimum and maximum projection of SLR as reported in Rahmstorf [2007].
Figure B7: Comparison of three 90% credible intervals generated in a perfect model experiment by varying assumptions about the observation properties. The gold envelope represents the 90% credible interval generated by assuming 200 independent and identical distributed (i.i.d.) residuals. The gray envelope is the 90% credible interval generated from an AR(1) process and accounts homoskedastic errors. The blue envelope denotes the 90% credible interval produced by accounting for heteroskedastic errors and the AR(1) process. The points are the synthesized observations and the red lines represent the generated observational errors.
Figure B8: Test of the impact of increasing the number of observations on the probability density function (parameter a, panel a & parameter b, panel b) and reliability diagram (panel c & d) in the simple linear test case. 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 (1-200) used in calibration.
Figure B9: Test of the impact of varying assumptions about the observation properties on the probability density function (parameter a, panel a & parameter b, panel b) and reliability diagram (panel c & d) in the simple linear test case. 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 lines and points represent calibration with different assumptions about the observation properties; gold as 200 i.i.d. residuals, gray as autocorrelated residuals, and blue as autocorrelated and heteroskedastic residuals.
Figure B10: Projections for global mean SLR in 2100 determined for the four methods presented as a probability density function (panel a), a cumulative density function (panel b), and a survival function (panel c). The horizontal dashed lines in the survival function (panel c) represent the reliability of 1:100 and 1:10,000 used in example studies to design SLR adaptation strategies [IWR, 2011; Houston, 2013].
**Table B1.** Prior distributions and calibrated minimum, median, and maximum estimates for model parameters.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Distribution</th>
<th>Prior</th>
<th>Rahmstorf (2007)</th>
<th>Bootstrap Posterior</th>
<th>MCMC Homoskedastic Posterior</th>
<th>MCMC Heteroskedastic Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLR sensitivity (a) [cm/yr/C]</td>
<td>Uniform</td>
<td>0 to 2</td>
<td>0.34</td>
<td>Min. 0.13</td>
<td>0.01</td>
<td>2.0x10^{-4}</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Med. 0.30</td>
<td>0.27</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Max. 0.46</td>
<td>0.64</td>
<td>1.36</td>
</tr>
<tr>
<td>Equilibrium temp. (T0) [°C]</td>
<td>Uniform</td>
<td>-3 to 2</td>
<td>-0.5</td>
<td>Min. -1.18</td>
<td>-3.0</td>
<td>-3.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Med. -0.56</td>
<td>-0.61</td>
<td>-0.81</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Max. -0.38</td>
<td>-0.22</td>
<td>0.26</td>
</tr>
<tr>
<td>Initial value in 1880 (H0) [cm]</td>
<td>Uniform</td>
<td>-16.8 to -12.5</td>
<td>-0.1464</td>
<td>Min. -16.44</td>
<td>-16.79</td>
<td>-16.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Max. -13.70</td>
<td>-12.49</td>
<td>-12.49</td>
</tr>
<tr>
<td>Standard deviation (σ_{AR1})</td>
<td>Uniform</td>
<td>0 to 1</td>
<td>NA</td>
<td>Min. 0.37</td>
<td>0.37</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Med. 0.48</td>
<td>0.49</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Max. 0.58</td>
<td>0.69</td>
<td>0.57</td>
</tr>
<tr>
<td>Lag-1 auto-correlation coefficient (ρ)</td>
<td>Uniform</td>
<td>-1 to 1</td>
<td>NA</td>
<td>Min. NA</td>
<td>0.37</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Med. 0.66</td>
<td>0.73</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Max. NA</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>