Evaluation of climate model simulations by means of statistical methods

Ekaterina Fetisova

Evaluation of climate model simulations by means of statistical methods

Ekaterina Fetisova

Licentiate thesis in Mathematical Statistics at Stockholm University to be publicly presented on Tuesday 17 November 2015 at 13.00 in room 306, building 6, Kräftriket, Roslagsvägen 101.

Typeset by LaTeX

Author e-mail: katarina@math.su.se

urn.kb.se/resolve?urn=urn:nbn:se:su:diva-122032

Printed in Sweden by Eprint AB, Stockholm 2015
Distributor: Department of Mathematics, Stockholm University

To the memory of my mother, Evgeniya

Abstract

Evaluation of climate models is a key issue within climate research. The statistical framework proposed by Sundberg et al., 2012, provides the theoretical underpinnings of methods for evaluation of climate model simulations by use of climate proxy data from the last millennium. In the present work, the statistical framework above is used to suggest several latent factor models of different complexity that can be used for estimating the amplitude of a forcing effect in a climate model by comparison with the observed/reconstructed climate. The performance of the models is evaluated and compared in a pseudo-proxy experiment, in which the true unobservable temperature series is replaced by selected realizations of a climate simulation model. For different levels of added noise, different conclusions can be drawn. However, for realistic noise levels, we find that the simplest model, the just-identified two-indicator one-factor model, denoted `j.i.FA(2,1)`, is a competitive alternative to models with more complicated structure. Moreover, we discover that the Fieller method of constructing confidence regions, associated with the `j.i.FA(2,1)`-model, outperforms the Wald confidence interval, which in most cases fails to provide sensible and interpretable conclusions about the climate model under consideration. Last but not least, the results indicate a good performance of the `j.i.FA(2,1)`-model even in the presence of heteroscedasticity.

Keywords: *Climate models*, *Climate proxy*, *Pseudo-proxy experiment*, *Factor analysis*, *the Wald confidence interval*, *the Fieller confidence set.*

Abbreviations

`j.i.ME`-model	- just-identified Measurement Error model
`j.i.FA(2,1)`-model	- just-identified Factor Analysis model with 2 indicators and 1 latent factor
`o.i.ME`-model	- overidentified Measurement Error model
`o.i.FA(2,1)`-model	- overidentified Factor Analysis model with 2 indicators and 1 latent factor
`j.i.FA(2,2)`-model	- just-identified Factor Analysis model with 2 indicators and 2 latent factors
`j.i.FA(3,2)`-model	- just-identified Factor Analysis model with 3 indicators and 2 latent factors
`o.i.FA(3,2)`-model	- overidentified Factor Analysis model with 3 indicators and 2 latent factors
`o.i.FA(3,1)`-model	- overidentified Factor Analysis model with 3 indicators and 1 latent factor

Acknowledgments

First of all, I am very grateful to all of my supervisors: Associate Professor Gudrun Brattström, Professor Rolf Sundberg (both at the department of Mathematics, division of Mathematical Statistics), and Associate Professor Anders Moberg at the department of Physical Geography, for their confidence in me and for supporting me continuously and patiently in my research. I also thank them for introducing me to the research group Griffin, in which all of them are involved. Meeting other researchers at Griffin meetings gave me a deeper and broader understanding of statistical and climatological issues. I thank all of the participants for rewarding presentations and for the friendly atmosphere at all our meetings. A special thanks goes to Alistair Hind and Qiong Zhang for their help with providing and understanding data.

Finally, I heartily thank my family in Sweden and in Russia, especially my husband, Khosro Lashgari, for all their affection and for encouraging me in my studies.

Contents

Abstract **i**

Abbreviations **iii**

Acknowledgments **v**

1 Introduction **9**

2 Theoretical background **16**
2.1 Basic statistical model and aim of the analysis 16
2.2 Decomposition of the total forcing effect 17
2.3 Models with one latent factor. Homoscedasticity 20
2.3.1 Approach 1: $\kappa = 1$ 20
2.3.2 Approach 2: $\mathrm{Var}(\xi_{\mathsf{f}}) = 1$ 26
2.3.3 Relation to other studies 33
2.4 Models with one latent factor. Heteroscedasticity. Approach 1: $\kappa = 1$ and Approach 2: $\mathrm{Var}(\xi_{\mathsf{f}}) = 1$ 36
2.5 Models with two latent factors. Homoscedasticity 41
2.5.1 Approach 1: $\kappa = 1$ 41
2.5.2 Approach 2: $\mathrm{Var}(\xi_{\mathsf{f}}) = 1$ 46
2.5.3 Extension of two-factor models. Approach 1: $\kappa = 1$ . . 48
2.5.4 Extension of two-factor models. Approach 2: $\mathrm{Var}(\xi_{\mathsf{f}}) = 1$ 49
2.6 Models with two latent factors. Heteroscedasticity 50
2.7 Usage of mean time series 51
2.8 Summary 52

3 Numerical experiment to compare the statistical models **54**
3.1 Description of the pseudo-proxy experiment 54
3.2 Description of data, its initial analysis and preliminaries . . . 56
3.3 Numerical results for data with zero proxy noise 62
3.4 Sensitivity to increasing noise 68
3.5 Estimation of parameters under heteroscedasticity 76

4 Conclusions **78**

5 References **80**

6 Appendix **83**

1 Introduction

Current trends in the climate with the increasing frequency and severity of extreme events such as heat waves, droughts, flooding events and storms makes the issue of sustainable development of our society one of the vital questions for governments and communities in all parts of the world. Although the concept of sustainable development, including such elements as economic growth, eradication of poverty, environmental protection, job creation, security, and justice (Victor et al., 2014) can have different goals in different countries, the joint achievement of these goals is closely related to the climate and its variations. While some climate changes can be beneficial for human and economial development, other can be disruptive for a sustainable future.

To understand and predict the future climate variability it is crucial to understand not only how the climate varied in the past and how it varies now but also the mechanisms behind the climate system variability. An important tool to help us understand how the climate system works is climate models. Prior to defining a climate model, some climatological notations and definitions need to be introduced, and we start with the definition of climate and the climate system structure (two main sources have been used throughout the whole introductory section: Goosse et al., 2010, and McGuffie&Henderson-Sellers, 2005).

Climate is traditionally defined as the description, in terms of the mean and variability over a 30-year reference period, of the relevant atmospheric variables (e.g. temperature, precipitation, winds). In a wider sense, it is the statistical description of the climate system. The *climate system* consists of five major components: *the atmosphere*, *the hydrosphere*, that is the water on and underneath the Earth's surface (ocean, seas, rivers, lakes, underground water), *the cryosphere*, that is the portion of the Earth's surface where water is in solid form (sea ice, lake and river ice, snow cover, glaciers, ice caps and ice sheets), *the land surface* and *the biosphere*. All these components are in turn components of the broader system, the Earth system, which also includes geological processes, such as plate tectonics, that can be of importance for climate on very long time scales of millions to hundreds of million years. Hence, the understanding of numerous processes, taking place in each component of the climate system, and possible interactions between them requires the understanding of factors that have triggered these processes.

Usually factors that influence the climate system fall into two separate categories: external factors and internal factors. Examples of external fac-

tors are changes in solar radiation or in the orbital position of the Earth. Internal factors, as indicated by their name, are factors internal to the climate system itself. Ocean and atmosphere circulation and their variations and mutual interactions are examples of processes that are clearly internal to the climate system. Moreover, they are of natural character. Another internal factor, inducing natural climate changes, is volcanism. On short time scales, volcanic eruptions affect climate during a few years after an eruption through the release of small particles and various chemical compounds several tenths of kilometers up in the atmosphere. These particles interact with incoming solar radiation and also affect cloud properties and thereby affect climate until they have been washed out by precipitation, but climate does not interact with the volcanism on these time scales.

Beside natural internal factors, there exist internal factors that are of anthropogenic character, i.e. causing human-induced changes. The most prominent example of anthropogenic climate influence is the ongoing release of carbon dioxide to the atmosphere, primarily by burning fossil fuels and cement production. Other examples of human influence on climate are the emissions of aerosols through various industrial and burning processes, changes in land-use and the depletion of stratospheric ozone through emissions of halocarbons.

As a matter of fact, it is sometimes difficult to draw a clear boundary between external and natural internal forcings. The distinction really depends upon the time- and space-scales considered. For instance, whether the human influence should be considered as an external or internal factor would depends on how one conceptualises the problem of current interest, but in many situations human climate influence is considered as an external factor, the same holds for the volcanism.

In order to compare the magnitude of the changes in different factors and to evaluate their effect on the climate, it is often convenient to analyze their impact on the radiative balance of the Earth. The net change in the Earth's radiative balance at the tropopause (incoming energy flux minus outgoing energy flux) caused by the change in a climate factor is called *a climate (radiative) forcing*. Radiative forcings are measured in Wm^{-2} and they may vary depending on spatial and temporal scale under consideration.

So in addition to being classified, depending on their origin, as external or internal, natural or anthropogenic, forcings can be negative or positive in comparison with a previous state. An example of a positive forcing is the increase in the atmospheric concentration of carbon dioxide since 1750, of which most is certainly due to human factors. The contribution from carbon dioxide alone is estimated to be $+1.68\ \mathrm{Wm}^{-2}$, with an uncertainty

of +1.33 to +2.03 Wm^{-2} (IPCC, 2013). The total forcing from all greenhouse gases is +3.00 Wm^{-2}, with uncertainty +2.22 to +3.78 Wm^{-2}, while the total anthropogenic radiative forcing for the year 2011 relative to 1750 has been estimated to be +2.29 Wm^{-2} on average across the globe, with an uncertainty lying in the range +1.13 to +3.33 Wm^{-2}. An example of a negative radiative forcing is the forcing associated with increased amounts of sulphate aerosols in the atmosphere, which can be both of natural (explosive volcanic eruptions) and anthropogenic (fossil fuel burning, in particular coal burning) nature. The main effect of sulphate aerosols is the scattering of a significant fraction of the incoming solar radiation back to space, which induces a local warming in the stratosphere and a cooling below, but they also affect clouds and thereby affect climate by changed cloud properties and cloud amounts. According to IPCC (2013), the current total radiative forcing from all kinds of aerosols in the atmosphere is negative: –0.9 Wm^{-2}, with uncertainty –1.9 to –0.1 Wm^{-2}.

Powerful tools to investigate the effect of changes on the climate system and to produce scenarios for future climate changes are climate models. Based on physical, biological and chemical principles, climate models can be defined as a system of partial differential equations that represents the processes in the climate system. In constructing a model of the climate system the following components are of importance:

1. *Radiation* - the way in which the input of solar radiation to the atmosphere or ocean and the emission of infrared radiation are handled, e.g. through absorption and scattering;

2. *Dynamics* - the movement of energy around the globe by winds and ocean currents and vertical movements (e.g. small-scale air turbulence and deep-water formation);

3. *Surface processes* - inclusion of the effects of sea and land ice, snow, vegetation and the resultant change in albedo[1], surface-atmosphere energy and moisture interchanges;

4. *Chemistry* - the chemical composition of the atmosphere and the interactions with other components (e.g. carbon exchanges between ocean, land and atmosphere);

5. *Resolution in both time and space* - the timestep of the model and the horizontal and verical scales resolved.

[1] From the Latin *albus*, meaning white. It is the reflected fraction of incident radiation.

In practice, it is impossible to construct a climate model that can completely represent all processes at the time scales they are associated with. Moreover, some processes are still not sufficiently known to include their detailed behaviour in models. Therefore, the concept of parametrization of processes is a key concept within climate modeling. The time-scale being modelled determines the relative importance of processes and in what way they should be parameterized. The simplest form is the null parameterization where a process, or group of processes, is ignored. By intentionally neglecting some processes it is possible to identify the role of a particular process clearly or to test a hypothesis. In addition, unnecessary computing time will not be spent on processes that can be represented in simpler form. Depending on the time-scale on which other more important processes (for a particular situation) have been modelled explicitly, a particular process can be fully prescribed in form of a fixed boundary condition or can evolve interactively, for example the topography of the ice sheet in a model designed to study climate variations on a longer time scale. Representations of external forcings in climate models are similarly handled. They can be either represented by their reconstructions or directly computed if a model includes a representation of a corresponding process. In sum, parameterizations are usually not valid for all possible conditions, so that there is inherent uncertainty in the results.

In addition to being characterized by the number of components/processes that are represented interactively, climate models can also be characterized by the complexity of the processes that are included. The wide range of climate models includes

- simple `Energy Balance Models` (EBMs). They are often zero- or one-dimensional models, typically predicting the surface (strictly the sea-level) temperature as a function of the energy balance of the Earth. But the way in which radiation is absorbed, transferred and re-emitted by the atmosphere is heavily simplified by means of parametrization of those processes;

- `Earth Models of Intermediate complexity` (EMICs) deal explicitly with surface processes and dynamics, often in a zonally averaged representation of the atmosphere and the ocean. They can be of varying degree of complexity and even be three-dimensional, where some particular components of the climate system may be described in great detail;

- `Coupled Climate Models`. They are complex fully coupled three-

dimensional models of the atmosphere and ocean incorporating other components such as the sea ice, the carbon cycle, ice sheet dynamics and even atmospheric chemistry. The core of these models is a General Circulation Model (GCM) that describes the three-dimensional atmosphere and ocean dynamics. Separate models for the other components of the climate system are coupled to the GCM through a model coupler. Such coupled models are often called Earth System Models (ESMs) or Coupled Global Climate Models (CGCMs).

Depending on the objective, one type of models could be selected. On the other hand, it is not unusual that the results from various types of models are combined in climate research. Enhanced computational and storage capacity of computers have led to the idea of 'ensemble runs' of the same model. In such experiments, the modellers let the external forcings be the same for all runs, but carefully perturb initial conditions for each model run, producing an ensemble set. Such experiments help place limits on the variation in climate. The availability of ensembles is also valuable from the statistical point of view because a simulation ensemble corresponds to a set of replicates in statistical terminology.

When a climate model is developed, it has to be tested to assess its quality and evaluate its performance. A first step is to ensure that the numerical model solves the equations of the physical model adequately. This procedure, often referred to as verification, only deals with the numerical resolution of the equations in the model, not with the agreement between the model and reality. It checks that no coding errors have been introduced into the program.

The next step is the validation process, i.e. determining whether the model accurately represents reality. To do this, the model results have to be compared with observations obtained under the same conditions. In particular, this implies that data input must be correctly specified to represent the observed situation. The agreement should be related to the intended use of the model. This could be done more or less intuitively by visually comparing maps or plots describing both the model results and the observations. Another way to compare is to define an appropriate metric, for example a simple root mean square (RMS) error:

$$RMS = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(T_{k,\mathrm{model}} - T_{k,\mathrm{obs}}\right)^2},$$

where k represents grid points for which observations are available, $k = 1, 2, \ldots, n$, $T_{k,\mathrm{model}}$ is the climate model variable of interest, for example the

model annual mean surface temperature at point k, and $T_{k,\mathrm{obs}}$ is then the observed annual mean surface temperature at point k. The RMS errors of different variables can be combined in various ways. It is also important that the model data-comparison should also take into account the errors or uncertainties in both the model results and the observations. Errors in the observations can be related to the precision of the instruments or to the way individual observations have been used to construct gridded data set. One may also treat the internal variability of the climate system as errors in this context.

An important stage in the development of climate models, and also in investigations aimed to understand properties of the real climate system, is a series of sensitivity tests. The behaviour of modelled climate systems is examined by altering one component, which enables to study the effect of this change on the model's climate. Usually sensitivity is described as a unit of response per unit change in a known forcing.

Because the modern instrumental climate record is very short compared to the geological history of the Earth, the available instrumental observations do not cover the full range of variability that a climate model should be able to represent. Therefore, many studies were devoted to comparison of climate model simulations with paleodata for different past climate situations (see examples in Texier et al, 1997; Brewer et al, 2007; Braconnot et al, 2012). The common feature of the methods applied is that they involve the observed output from the real world climate system, as recorded in the climate proxy data, and the observed output from the simulated climate system.

Recently a new statistical framework for evaluation of climate model simulations against a diverse set of climate proxy series has been developed by Sundberg et al., 2012 (hereafter referred to as SUN12). This framework was specifically developed to suit the comparison of simulations and proxy data for the relatively recent past of about one millennium or so, when a large number of climate proxy data series having annual resolution exist and when many simulations with different coupled global climate system models have already been performed (PAGES2k-PMIP3 group, 2015).

The distinctive feature of this framework is that it treats the real climate system and the simulated climate in terms of *unobservable* temperature changes caused by external and internal factors. This gives an opportunity to develop new methods of evaluating climate models in addition to those that are already widely applied. Indeed, apart from the correlation and distance test-statistics, developed in SUN12, the framework provides a theoretical basis for evaluation of climate model simulations by compar-

ing the amplitude of an unobservable simulated forcing effect, caused by a particular (reconstructed) forcing that constitutes a forcing history of the climate model under consideration, with the amplitude of an unobservable real-world forcing effect caused by the real-world counterpart of the reconstructed forcing. Agreement in the amplitudes is interpreted then as the agreement between the real-world forcing and its reconstruction.

The first step in this direction was recently taken by Tingley et al. (2015) by analyzing a certain type of the measurement error (ME) model, formulated on the basis of the statistical framework of SUN12, by Bayesian methods. Since our own analysis is also based on this statistical framework, a comparison of the suggested methods is of interest. Without aiming to perform a detailed evaluation of the analysis carried by Tingley et al. (2015), we will present a brief theoretical comparative discussion in Sec. 2.3.3.

It should be remarked that the concept of latent variables is not new within climate research. A prominent example of its application is the optimal fingerprinting framework used in detection and attribution (D&A) studies (Mitchell et al, 2001; Hegerl et al, 2007, 2011), seeking to identify the latent forced response in temperature reconstructions. In Sec. 2.3.3, we also elucidate the link between our methods and one of the methods used in the D&A studies.

At this point, we may move on to the theoretical part of our analysis by starting with the description of the statistical framework formulated in SUN12. As in SUN12, the entire discussion here is made bearing in mind the properties of data being available for the last millennium or so. Nevertheless, the statistical models discussed here are general and should also be valid for other time periods extending further back. However, whether they have any practical value or not, depends on whether the available climate model simulation and climate proxy data allow them to be used or not.

2 Theoretical background

2.1 Basic statistical model and aim of the analysis

The statistical framework in SUN12 includes the formulation of the following model for data under *forced* climate model simulations:

Basic Statistical Model

$$\begin{aligned} x_t &= \mu_x + \alpha \cdot \xi_t + \delta_t \\ \tau_t &= \mu_\tau \; + \quad \xi_t + \eta_t \\ y_t &= \tau_t + \theta_t \\ z_t &\approx \tau_t + \epsilon_t, \end{aligned} \tag{2.1}$$

where $t = 1, 2, \ldots, n$, and

x_t - a simulated temperature value, generated by a climate model, for the region of interest and time point t.

τ_t - a true temperature, corresponding to x_t. The true temperature is an unobserved variable, i.e. latent variable.

y_t - a measured temperature, averaged over the same region and time unit.

z_t - a properly calibrated climate proxy, which serves as a surrogate for the true temperature τ_t (the calibration method can be found in SUN12).

μ_x, μ_τ - the mean values over time, around which x, τ, y and a calibrated proxy z vary.

ξ_t - the true effect of a specific type of forcing that has influenced the true temperature τ_t. The forcing can be either of a single type (e.g. only volcanic forcing) or a combination of several forcings (e.g. volcanic and solar forcing).

$\alpha\xi_t$ - represents the unknown variability in x that can be linearly explained by the true forcing effect. A correct representation of the forcing effect in the climate model corresponds to $\alpha = 1$, whereas an unforced climate model has $\alpha = 0$.

δ_t - represents internal noise variability in the simulations and any variability in the simulations unrelated to the true forcing effects.

η_t - denotes the residual variation in true temperature that cannot be statistically explained by the particular forcing under consideration.

θ_t - denotes the measurement error in the observed temperature y_t, uncorrelated with τ_t.

ϵ_t - represents the residual variation in z, uncorrelated with τ_t.

Quantities δ, η, θ and ϵ are regarded as mutually uncorrelated random variables, with mean values zero and variances σ_δ^2, σ_η^2, σ_θ^2 and σ_ϵ^2. The first three variances are assumed to be constant over time, while the last one may vary with time.

The aim of the present analysis is the estimation of the parameter α. To avoid any ambiguity in its interpretation, the viewpoint suggested by model (2.1) has been adopted throughout the whole work, namely an estimate of α provides a measure of the amplitude of a simulated forcing effect associated with a reconstructed external forcing used for generation of $\{x_t\}$ by a climate model under consideration.

2.2 Decomposition of the total forcing effect

We start approaching the aim by examining first the structure of the expression for the true temperature τ. Let us assume for a moment that τ_t is observable. According to the model above, the *mean-centered* τ_t is given by

$$\tau_t = \xi_{\mathsf{f}\,t} + \eta_t \tag{2.2.1}$$

where the index f emphasizes the fact that the forcing effect ξ is due to a particular forcing f. In reality, however, the true temperature is affected by all external forcings simultaneously, implying the following decomposition of τ:

$$\tau_t = \xi_{\mathsf{total}\,t} + \eta_{\mathsf{internal}\,t}, \tag{2.2.2}$$

where ξ_{total} is the total forcing effect and η_{internal} denotes temperature changes due to factors that are internal to the climate system itself. It is assumed that $\xi_{\mathsf{total}\,t}$ and $\eta_{\mathsf{internal}\,t}$ are uncorrelated for each time point t.

To rewrite (2.2.2) in the form of (2.2.1), we need first to isolate ξ_{f} from

the total forcing effect ξ_{total}. A possible way to do it is to let $\boldsymbol{\xi}_{\mathsf{total}}$ be orthogonally projected on $\boldsymbol{\xi}_{\mathsf{f}}$. Denoting the orthogonal projection by $\kappa \cdot \boldsymbol{\xi}_{\mathsf{f}}$ and the orthogonal complement by $\boldsymbol{\xi}_{\mathsf{total}\perp\mathsf{f}}$, the total forcing effect may be represented in the following way (see Figure 1 for a graphical illustration):

$$\boldsymbol{\xi}_{\mathsf{total}} = \kappa \cdot \boldsymbol{\xi}_{\mathsf{f}} + \boldsymbol{\xi}_{\mathsf{total}\perp\mathsf{f}}. \tag{2.2.3}$$

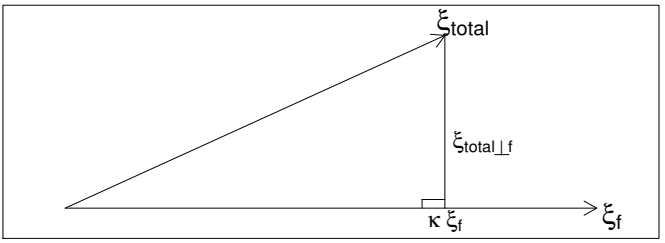

Figure 1. Schematical description of decomposing the total forcing effect into two parts.

Inserting (2.2.3) into (2.2.2) leads to

$$\tau_t = \kappa \cdot \xi_{\mathsf{f}\,t} \;+\; \xi_{\mathsf{total}\perp\mathsf{f}\;t} + \eta_{\mathsf{internal}\;t}. \tag{2.2.4}$$

Due to the orthogonal decomposition of the total forcing effect, all components in (2.2.4), i.e. $\kappa \cdot \xi_{\mathsf{f}\,t}$, $\xi_{\mathsf{total}\perp\mathsf{f}\;t}$ and $\eta_{\mathsf{internal}\;t}$, are mutually uncorrelated. This allows us to consider $\xi_{\mathsf{total}\perp\mathsf{f}\;t}$ as a part of the residual variation in τ that cannot be statistically explained by the particular forcing f. According to the basic statistical model, this residual variation is denoted by η_t. That is, we may write

$$\tau_t = \kappa \cdot \xi_{\mathsf{f}\,t} \;+\; \underbrace{\eta_t}_{=\xi_{\mathsf{total}\perp\mathsf{f}\;t}+\eta_{\mathsf{internal}\;t}}. \tag{2.2.5}$$

It remains to explain the reason behind setting the coefficient κ to 1 in the basic statistical model. To this end, let Eq.(2.2.5) together with the expression for *mean-centered* simulated temperature, $x_{\mathsf{f}\,t}$, form the following equation system:

$$\begin{cases} x_{\mathsf{f}\,t} = \alpha_{\mathsf{f}} \cdot \xi_{\mathsf{f}\,t} \;+\; \delta_{\mathsf{f}\,t} \\ \tau_t \;\; = \kappa \cdot \xi_{\mathsf{f}\,t} \;+\; \eta_t \end{cases} \tag{2.2.6}$$

This is a factor model with one latent factor, ξ_{f} , which is assumed to be responsible for the correlation among the two observed variables, x and τ. In the terminology of factor analysis observed variables are called *indicators* or *manifest* variables. The coefficients α_{f} and κ are the model loadings, while the errors δ_{f} and η_t are called *specific* factors because they are specific to the particular indicator they are associated with. Specific-factor variables are assumed to be uncorrelated with latent factors, implying that if we eliminate the effect of latent factors on indicators, indicators become mutually uncorrelated. Moreover, as required by most theoretical results, specific-factor variables are assumed to be identically and independently distributed (abbr. i.i.d.). Although this assumption can hardly be met in real-world climate data, the whole theoretical discussion in this section will be based on the assumption of independent observations. The issue of autocorrelation will be discussed in Sec. 3. The factor loading α_{f} in (2.2.6) is estimable, or equivalently identified, if:

1. The factor loading κ is fixed to one or

2. The variance of ξ_{f} is fixed to one.

As we see, in the basic statistical model the first identification approach, i.e. $\kappa = 1$, has been applied.

At this point, we have to admit that regardless of approach, we need observed data to estimate the parameter of interest. A natural way to construct such data is to concatenate observed temperatures and a calibrated proxies in the following way: for the period when y is observed, τ_t is replaced by the measured y_i, while outside of this period by a calibrated proxy z_t. Based on the calibration method described in SUN12, the (complete) climate record, hereafter denoted by v, is given by

$$v_t = \begin{cases} z_t \approx \tau_t + \epsilon_t & t \in \text{the period when only } z \text{ is available,} \\ & \text{the so-called reconstruction period} \\ \\ y_t = \tau_t + \theta_t & t \in \text{the period when both } y \text{ and } z \text{ are available,} \\ & \text{the so-called calibration period.} \end{cases} \tag{2.2.7}$$

Typically in applications, the variance of noise in the proxies, σ^2_ϵ, is much larger than σ^2_θ. In addition, it might substantially vary within the reconstruction period. Therefore, replacing the true temperature by the climate record $\{v_t\}$, we are faced with the issue of time-varying variances, known as *heteroscedasticity*. To understand how heteroscedasticity can be

taken into account, let us first derive an estimator of α_{f} in the absence of it, in other words in the presence of homoscedasticity.

2.3 Models with one latent factor. Homoscedasticity

2.3.1 Approach 1: $\kappa = 1$

Consider model (2.2.6). Setting κ to 1 and replacing τ by calibrated proxies with a time-constant precision, we obtain:

$$\begin{cases} x_{\mathsf{f}\,t} &= \alpha_{\mathsf{f}} \cdot \xi_{\mathsf{f}\,t} + \delta_{\mathsf{f}\,t} \\ z_t &= \xi_{\mathsf{f}\,t} + \underbrace{\nu_t}_{=\eta_{\,t}+\epsilon_t} \end{cases} \tag{2.3.1}$$

where ν_t has a constant variance $\sigma^2_\nu = \sigma^2_\eta + \sigma^2_\epsilon = \sigma^2_{\xi_{\mathsf{total}\perp\mathsf{f}}} + \sigma^2_{\eta_{\mathsf{internal}}} + \sigma^2_\epsilon$. Model (2.3.1) is known as a *measurement error* (ME) model (for its basic definition see Cheng, 1999, Sec. 1.1 or Fuller, 1987, Sec. 1.1.1), where the unobservable variable ξ_{f} might be either fixed or random. Models with ξ_{f} regarded as fixed are called *functional* models, while models with ξ_{f} regarded as random are called *structural* models. In particular, under normality assumption:

`Structural model:` $\xi_{\mathsf{f}\,t} \overset{\text{i.i.d}}{\sim} \mathrm{N}(0, \sigma^2_{\xi_{\mathsf{f}}}), \qquad (\delta_{\mathsf{f}\,t}, \nu_t)' \overset{\text{i.i.d}}{\sim} \mathrm{N}\left(\mathbf{0}, \mathrm{diag}(\sigma^2_{\delta_{\mathsf{f}}}, \sigma^2_\nu)\right)$

`Functional model:` $\frac{1}{n}\sum_t \xi_{\mathsf{f}\,t} = 0, \quad \sigma^2_{\xi_{\mathsf{f}}} \equiv \frac{1}{n}\sum_t \xi^2_{\mathsf{f}\,t} > 0, \qquad (\delta_{\mathsf{f}\,t}, \nu_t)' \overset{\text{i.i.d}}{\sim} \mathrm{N}\left(\mathbf{0}, \mathrm{diag}(\sigma^2_{\delta_{\mathsf{f}}}, \sigma^2_\nu)\right).$

Despite the restriction $\kappa = 1$, it is still not possible to estimate the parameter α_{f}. More restrictions on the model parameters need to be imposed in order to make α_{f} identified, i.e. estimable. A definition of identified parameters and models is as follows.

Definition 1. The parameter θ_i, where θ_i is the ith element of the vector of the model parameters $\boldsymbol{\theta}$, is identified if no two values of $\boldsymbol{\theta}$, belonging to the space of possible parameter values Θ, for which θ_i differ, lead to the same sampling distribution of the indicators. The model is identified if and only if every element of $\boldsymbol{\theta}$ is identified.

The parameter α_{f} is identified if one of the following conditions is satisfied (see Cheng, 1999, sec. 1.2.1):

(a) $\sigma^2_{\delta_{\mathsf{f}}}$ is known;

(b) σ^2_{ν} is known;

(c) both of the error variances, $\sigma^2_{\delta_{\mathsf{f}}}$, and σ^2_{ν}, are known; (2.3.2)

(d) the ratio of the error variances, $\sigma^2_{\delta_{\mathsf{f}}}/\sigma^2_{\nu}$, is known;

(e) the ξ_{f} reliability ratio, $\sigma^2_{\xi_{\mathsf{f}}}/\sigma^2_{z}$ where $\sigma^2_{z} = \sigma^2_{\xi_{\mathsf{f}}} + \sigma^2_{\nu}$, is known.

In practice, neither the ratios nor the individual variances can be known. Perhaps the most realistic situation is that one of the error variances is estimated by an independent estimator.

The internal variability of a climate model, $\sigma^2_{\delta_{\mathsf{f}}}$, can be estimated by means of replicates of the x_{f} -climate model[2], provided that such replicates are available. At least two sequences are needed. More replicates will lead to a more precise estimate of $\sigma^2_{\delta_{\mathsf{f}}}$. Letting k denote the number of replicates of x_{f}, the estimate is given by

$$\widehat{\sigma}^2_{\delta_{\mathsf{f}}} = \frac{\sum_{t=1}^{n}\sum_{i=1}^{k}(x_{\mathsf{f},\ \mathsf{repl.i}\ t} - \bar{x}_{\mathsf{f}\ .t})^2}{n(k-1)}, \tag{2.3.3}$$

where $\bar{x}_{\mathsf{f}\ .t}$ is the average of k replicates at time point t. If there are exactly two replicates, $\widehat{\sigma}^2_{\delta_{\mathsf{f}}}$ is half the sample variance of the difference sequence $\{x_{\mathsf{f}\ i1} - x_{\mathsf{f}\ i2}\}$. Note that this estimator is unbiased even in the presence of autocorrelation in $\{x_{\mathsf{f},\ \mathsf{repl.i}\ t}\}$, which is due to the independence of replicates.

Regarding the variance of ν, we have to remember that it consists of several variances, $\sigma^2_{\nu} = \sigma^2_{\epsilon} + \sigma^2_{\eta} = \sigma^2_{\epsilon} + \sigma^2_{\xi_{\mathsf{total}\perp\mathsf{f}}} + \sigma^2_{\eta_{\mathsf{internal}}}$. As argued in SUN12, the variance σ^2_{ϵ} is in principle estimable, and the variance of η_{internal} can at least be roughly approximated[3]. However, it seems impossible to determine an appropriate source for the estimation of $\sigma^2_{\xi_{\mathsf{total}\perp\mathsf{f}}}$. Recall that $\xi_{\mathsf{total}\perp\mathsf{f}}$ is in effect an orthogonal complement, whose variability depends on what particular forcing has been isolated from ξ_{total}. Therefore, we can conclude that the most appropriate and realistic identifiability condition is (2.3.2a), i.e. $\sigma^2_{\delta_{\mathsf{f}}}$ known. Despite the fact that this variance is in effect

[2]Replicates of a climate model are simulations with identical reconstructed forcing but with different initial conditions. In the terminology of climate modeling, replicates are called *members in a simulation ensemble*.

[3]see SUN12, p. 1345 for the proposed estimation methods and sources. Note that at this page $\sigma^2_{\eta_{\mathsf{internal}}}$ is referred to as σ^2_{η}.

estimated, we henceforth will refer to parameters whose values are known before a statistical model is fitted to data as *known parameters*. The ME model in (2.3.1) has two known parameters: the loading κ and the variance of $\delta_{\mathsf{f}\,t}$.

Under these assumptions, the model parameters are: α_{f}, $\sigma^2_{\xi_{\mathsf{f}}}$ and σ^2_{ν}. To determine their estimates, consider the covariance-variance matrix of the observed variables, where each nonduplicated (or unique) element is expressed as a function of the model parameters:

$$\begin{aligned} \sigma^2_{x_{\mathsf{f}}} &= \alpha^2_{\mathsf{f}} \cdot \sigma^2_{\xi_{\mathsf{f}}} + \sigma^2_{\delta_{\mathsf{f}}} \\ \sigma_{xz} &= \alpha_{\mathsf{f}} \cdot \sigma^2_{\xi_{\mathsf{f}}} \\ \sigma^2_{z} &= \sigma^2_{\xi_{\mathsf{f}}} + \sigma^2_{\nu}. \end{aligned} \tag{2.3.4}$$

Since there are three unique equations in three unknowns, there is one and only one way to solve for the unknowns. From (2.3.4) follows that the model parameters are uniquely determined by the following equations:

$$\begin{aligned} \alpha_{\mathsf{f}} &= \frac{\sigma^2_{x_{\mathsf{f}}} - \sigma^2_{\delta_{\mathsf{f}}}}{\sigma_{x_{\mathsf{f}}z}} \\ \sigma^2_{\xi_{\mathsf{f}}} &= \sigma_{x_{\mathsf{f}}z}/\alpha_{\mathsf{f}} = (\sigma_{x_{\mathsf{f}}z})^2/(\sigma^2_{x_{\mathsf{f}}} - \sigma^2_{\delta_{\mathsf{f}}}) \\ \sigma^2_{\nu} &= \sigma^2_{z} - \sigma^2_{\xi_{\mathsf{f}}} = \sigma^2_{z} - (\sigma_{x_{\mathsf{f}}z})^2/(\sigma^2_{x_{\mathsf{f}}} - \sigma^2_{\delta_{\mathsf{f}}}), \end{aligned} \tag{2.3.5}$$

Replacing the population variances and covariance of the indicators by their maximum likelihood (ML) estimates, $s^2_{x_{\mathsf{f}}}$, $s_{x_{\mathsf{f}}z}$ and s^2_z [4], the ML estimators of the model parameters can be obtained:

$$\begin{aligned} \widehat{\alpha}_{\mathsf{f}} &= \frac{s^2_{x_{\mathsf{f}}} - \sigma^2_{\delta_{\mathsf{f}}}}{s_{x_{\mathsf{f}}z}} \\ \widehat{\sigma}^2_{\xi_{\mathsf{f}}} &= s_{x_{\mathsf{f}}z}/\widehat{\alpha}_{\mathsf{f}} = (s_{x_{\mathsf{f}}z})^2/(s^2_{x_{\mathsf{f}}} - \sigma^2_{\delta_{\mathsf{f}}}) \\ \widehat{\sigma}^2_{\nu} &= s^2_{z} - \widehat{\sigma}^2_{\xi_{\mathsf{f}}}, \end{aligned} \tag{2.3.6}$$

[4] The ML estimates for the population variances and covariances in the bivariate normal distribution are $s^2_{x_{\mathsf{f}}} = \sum(x_{\mathsf{f}t} - \bar{x}_{\mathsf{f}})^2/n$, where $\bar{x}_{\mathsf{f}} = \sum x_{\mathsf{f}t}/n$, etc.. For small samples, it seems reasonable to use unbiased estimates for $\sigma^2_{x_{\mathsf{f}}}$, $\sigma_{x_{\mathsf{f}}z}$, and σ^2_z that differ from the ML estimates by a factor $n/(n-1)$. For large samples, however, both types of the estimates can be used because the difference between them is negligible.

provided three side conditions are fulfilled: (1) $s_{x_\mathsf{f} z} \neq 0$, (2) $s^2_{x_\mathsf{f}} > \sigma^2_{\delta_\mathsf{f}}$, and (3) $s^2_z - (s_{x_\mathsf{f} z})^2/(s^2_{x_\mathsf{f}} - \sigma^2_{\delta_\mathsf{f}}) \geq 0$.

The side conditions are motivated by the requirement that $\widehat{\alpha}_\mathsf{f}$ must be finite and the variance estimates, $\widehat{\sigma}^2_{\xi_\mathsf{f}}$, and $\widehat{\sigma}^2_\nu$, must be nonnegative. The first two assumptions indicate that $\sigma^2_{\xi_\mathsf{f}} > 0$. The third side assumption ensures that $\sigma^2_\nu \geq 0$. If this assumption is not satisfied, the solution for σ^2_ν is an inadmissible solution termed Heywood case. The usual interpretation of a Heywood case is that the corresponding true variance is small and estimated as zero. Having got an indication that $\sigma^2_\nu \approx 0$, the ME model simplifies to the ordinary regression model. The associated estimator is

$$\widehat{\alpha}_\mathsf{f} = \frac{s_{x_\mathsf{f} z}}{s^2_z} \tag{2.3.7}$$

As discussed by Moberg and Sundberg (1978), the estimators of α_f in (2.3.6) and (2.3.7) are the upper and lower bounds, respectively, for $\widehat{\alpha}_\mathsf{f}$ in the normal functional model provided $\sigma^2_{\delta_\mathsf{f}}$ is known and $s_{x_\mathsf{f} z} > 0$. However, in the climatological context, the situation with $\sigma^2_\nu = 0$ is senseless since it means that all incoming variances, including σ^2_ϵ, are zero. This contradicts our knowledge about the properties of proxies, whose non-climatic component ϵ is assumed to constitute the largest part of the observed proxy. It makes the use of estimator (2.3.7) unmotivated, except as an estimated lower bound.

An important remark about estimator (2.3.6) is that it is a ratio of two random variables. Such ratios are typically biased estimators of the ratio of the expectations. Nevertheless, the estimator is consistent regardless of whether ξ_t is random or fixed. By a consistent estimator we mean the following:

Definition 2. An estimator T_n, defined for every n, is *consistent* as an estimator of a parameter θ if for any $a > 0$,

$$P\left(|T_n - \theta| > a\right) \to 0 \qquad \text{as } n \to \infty.$$

If θ is a vector then $T_n - \theta$ is replaced by $||T_n - \theta||$.

In other words, we say an estimate is consistent if the estimate converges in probability to the true parameter, symbolically expressed as $T_n \xrightarrow{p} \theta$ as $n \to \infty$.

To see the consistency of $\widehat{\alpha}_\mathsf{f}$ in (2.3.6) we use the fact that the sample variance and covariance converge in probability to their expectations, that

is to the population variance and covariance. From that we get:

$$\widehat{\alpha}_{\mathsf{f}} = \frac{s^2_{x_{\mathsf{f}}} - \sigma^2_{\delta_{\mathsf{f}}}}{s_{x_{\mathsf{f}}z}} \quad \xrightarrow{p} \quad \frac{\sigma^2_{x_{\mathsf{f}}} - \sigma^2_{\delta_{\mathsf{f}}}}{\sigma_{x_{\mathsf{f}}z}} = \frac{(\alpha^2_{\mathsf{f}} \cdot \sigma^2_{\xi_{\mathsf{f}}}) + \sigma^2_{\delta_{\mathsf{f}}}) - \sigma^2_{\delta_{\mathsf{f}}}}{\alpha_{\mathsf{f}} \cdot \sigma^2_{\xi_{\mathsf{f}}}} = \alpha_{\mathsf{f}}. \tag{2.3.8}$$

Unfortunately, consistency, which is an asymptotic requirement, does not guarantee good properties of the estimator for finite sample size. This is due to the fact that for fixed n the estimator has infinite mean and infinite variance. On the other hand, it possesses an asymptotic distribution with both finite mean and variance. According to Cheng (1999, Sec. 2.1.3), the asymptotic (limiting) distribution of $\sqrt{n}(\widehat{\alpha}_{\mathsf{f}} - \alpha_{\mathsf{f}})$ is a normal distribution with zero mean and the variance equal to

$$\Gamma_{\alpha_{\mathsf{f}}} = \underbrace{\frac{1}{(\sigma^2_{\xi})^2 \cdot \alpha^2_{\mathsf{f}}}}_{=(\sigma_{xz})^2} \cdot \Big(\big(\alpha^2_{\mathsf{f}} \cdot \sigma^2_{\xi} + \sigma^2_{\delta_{\mathsf{f}}}\big) \cdot \big(\sigma^2_{\delta_{\mathsf{f}}} + \alpha^2_{\mathsf{f}} \cdot \sigma^2_{\nu}\big) + \sigma^4_{\delta_{\mathsf{f}}} \Big), \tag{2.3.9}$$

It is worth remarking that the variance in (2.3.9) is obtained in accordance with the so called delta method that, in its essence, consists in expanding an estimator in a first-order Taylor series. Applied to the ME model, this approach does not require ξ_{f} to be normally distributed, only that $(\delta_{\mathsf{f}}, \nu)$ be normal.

By replacing the unknown parameters in $\Gamma_{\alpha_{\mathsf{f}}}$ by their consistent estimates, we obtain an estimator of the variance of the limiting distribution of $\widehat{\alpha}_{\mathsf{f}}$:

$$\widehat{\mathrm{Var}}(\widehat{\alpha}_{\mathsf{f}}) = \frac{1}{n \cdot (s_{x_{\mathsf{f}}z})^2} \cdot \big(s^2_{x_{\mathsf{f}}} \cdot \big(\sigma^2_{\delta_{\mathsf{f}}} + \widehat{\alpha}^2_{\mathsf{f}} \cdot \widehat{\sigma}^2_{\nu}\big) + \sigma^4_{\delta_{\mathsf{f}}}\big)\,. \tag{2.3.10}$$

Because $n\widehat{\mathrm{Var}}(\widehat{\alpha}_{\mathsf{f}})$ is a consistent estimator of (2.3.9), it follows that for large samples

$$T = \frac{\widehat{\alpha}_{\mathsf{f}} - \alpha_{\mathsf{f}}}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\alpha}_{\mathsf{f}})}} \overset{\text{approx}}{\sim} \mathrm{N}(0,1). \tag{2.3.11}$$

Consequently, we may construct an approximate $100(1-p)\%$ confidence interval for α_{f}:

$$\widehat{\alpha}_{\mathsf{f}} \pm z_{p/2} \cdot \sqrt{\widehat{\mathsf{Var}}(\widehat{\alpha}_{\mathsf{f}})}, \tag{2.3.12}$$

where $z_{p/2}$ is the $100(1-p/2)$ percentile of the standard normal distribution. The confidence interval in (2.3.12) is known as the Wald confidence interval.

As follows from (2.3.6) and (2.3.10), a sufficiently large (in absolut value)

covariance between the simulated temperature and the observations is an important premise for obtaining a reasonable estimate of α_{f} and a reasonable confidence interval for it[5]. It suggests to test whether this covariance (or equivalently correlation between x_{f} and z) is statistically significantly different from zero before α_{f} is estimated. To this end, the correlation U_R test statistic, developed in SUN12 (its definition is also given in Appendix here), can be used.

In connection with using this test statistic, Moberg et. al (2015) pointed out that high correlation does not mean that the amplitude of the forcing effect is of the right size in the climate model. Analogously, we may say that low correlation does not necessarily mean that the amplitude of a forcing effect in a climate model simulation is of the wrong size. Indeed, based on the model's reproduced covariance between x_{f} and z, given by $\alpha_{\mathsf{f}} \cdot \sigma^2_{\xi_{\mathsf{f}}}$, low covariance can arise when α_{f} is arbitrarily small, or when $\sigma^2_{\xi_{\mathsf{f}}}$ is arbitrarily small, or when both of them are arbitrarily small. In all three cases, the ME model is close to an underidentifiable model, which makes the estimation procedure highly unstable.

Under the condition $\sigma^2_{\xi_{\mathsf{f}}} \approx 0$, arising when a true forcing f does not generate substantial temperature changes at the temporal and spatial scales analyzed, the correct value of α_{f} can be 1, but a small variability of ξ_{f} makes the estimation of α_{f} difficult, precisely as under ordinary linear regression where the explanatory variable does not vary much. Furthermore, when $\sigma^2_{\xi_{\mathsf{f}}} \approx 0$ the performance of the approximations used for deriving the asymptotic distribution of $\widehat{\alpha}_{\mathsf{f}}$ will not perform well and the confidence level of the confidence interval in (2.3.12) will decrease. As shown by Gleser and Hwang (1987), any confidence set for the slope in ME models of finite expected length must have confidence (confidence level) $1 - p = 0$, where the confidence level of a confidence set is defined to be the infimum of coverage probability over the parameter space. Because the Wald confidence interval in (2.3.12) has always finite length it has zero confidence level for any fixed n by virtue of Gleser and Hwang's theorem. Moreover, as n tends to infinity, the confidence level remains zero and does not tend to the nominal confidence level, $1 - p$.

Fortunately, this result does not mean that any given data set definitely suffer from the zero confidence level effect. One needs $\sigma^2_{\xi_{\mathsf{f}}}$ to be sufficiently

[5]In this analysis, an unreasonable confidence set is defined as a set containing both 1 and 0. From the climatological point of view, such confidence sets are completely useless because they allow two mutually exclusive types of interpretation: the event $\alpha_{\mathsf{f}} = 1$ means that the amplitude of a forcing effect in a climate model is correctly represented, while $\alpha_{\mathsf{f}} = 0$ means that a forcing is not represented at all.

large. However, it is not easy to test $H_0 : \sigma^2_{\xi_\mathsf{f}} = 0$ under the ME model under the assumption that $\sigma^2_{\delta_\mathsf{f}}$ is known. Gleser (1987) studied the confidence interval for the slope for the ME model when the ratio of the error variances is known. It was found that problems might arise if the signal to noise ratio, defined as

$$SNR = \sigma^2_{\xi_\mathsf{f}} / \sigma^2_\nu, \tag{2.3.13}$$

is less than 1. As assumed by Cheng (1999, p. 61), it is likely that Gleser's results can be extended to other assumptions despite the fact that Gleser's method of proof no longer works. However, keeping in mind the properties of real-world temperature proxies, it appears unrealistic, even for the ME model studied by Gleser, that the criterion $SNR \geq 1$ will be satisfied. It is well-known that climate proxies suffer from a large non-climatic noise that dominates the climate signal. Hence, the SNR can hardly aid in testing the hypothesis that $\sigma^2_{\xi_\mathsf{f}} = 0$ when climate data are analyzed.

Since the knowledge about the (unknown) variability of the forced component may contribute to a more comprehensive interpretation of the properties of a climate model simulation under consideration, we will investigate the properties of the model obtained under *Approach 2*, under which the variance of the forcing effect is known. More precisely, it is known to be 1.

2.3.2 Approach 2: $\mathbf{Var}(\xi_\mathsf{f}) = 1$

Let z, the proxy with a constant noise variance, be a substitute for the true unobservable temperature τ in model (2.7):

$$\begin{cases} x_{\mathsf{f}\,t} &= \alpha_\mathsf{f} \cdot \xi_{\mathsf{f}\,t} + \delta_{\mathsf{f}\,t} \\ z_t &= \kappa \cdot \xi_{\mathsf{f}\,t} + \underbrace{\nu_t}_{=\xi_{\mathsf{total}\perp\mathsf{f}\,t} + \eta_{\mathsf{internal}\,t} + \epsilon_t} \end{cases}. \tag{2.3.14}$$

As already mentioned, this is a two-indicator one-factor model, abbr. FA(2,1)-model. The distributional assumptions of $\xi_{\mathsf{f}\,t}$, $\delta_{\mathsf{f}\,t}$ and ν_t are the same as under the ME model in (2.3.1) with an additional restriction that $\sigma^2_{\xi_\mathsf{f}}$ is equal to 1. It implies that the unique equations of the associated variance-covariance matrix of the indicators expressed in terms of the model parameters are

$$\begin{aligned} \sigma^2_{x_\mathsf{f}} &= \alpha^2_\mathsf{f} + \sigma^2_{\delta_\mathsf{f}} \\ \sigma_{x_\mathsf{f}\,z} &= \alpha_\mathsf{f} \cdot \kappa \\ \sigma^2_z &= \kappa^2 + \sigma^2_\nu. \end{aligned} \tag{2.3.15}$$

The model parameteres are[6]: α_{f}, κ and σ^2_ν. Note that under the ME model we need to estimate α_{f}, $\sigma^2_{\xi_{\mathsf{f}}}$ and σ^2_ν. One might think that these two models are different, but they are by definition identical in the sense that they have the same estimator of the amplitude of a forcing effect in a climate model. To see it we rewrite (2.3.15) in the form identical with the corresponding matrix for the ME model:

$$\begin{array}{rcccc} & \underline{\textit{under the ME model in (2.3.1)}} & & \underline{\textit{under the FA(2,1)-model in (2.3.14)}} \\ \sigma^2_{x_{\mathsf{f}}} = & \alpha^2_{\mathsf{f}} \cdot \sigma^2_{\xi_{\mathsf{f}}} + \sigma^2_{\delta_{\mathsf{f}}} & = & \left(\frac{\alpha_{\mathsf{f}}}{\kappa}\right)^2 \cdot \kappa^2 + \sigma^2_{\delta_{\mathsf{f}}} \\ \sigma_{x_{\mathsf{f}}\, z} = & \alpha_{\mathsf{f}} \cdot \sigma^2_{\xi_{\mathsf{f}}} & = & \left(\frac{\alpha_{\mathsf{f}}}{\kappa}\right) \cdot \kappa^2 \\ \sigma^2_z = & \sigma^2_{\xi_{\mathsf{f}}} + \sigma^2_\nu & = & \kappa^2 + \sigma^2_\nu. \end{array} \tag{2.3.16}$$

From (2.3.16) it follows that the estimator of α_{f} under the ME model is the same as the estimator of $\alpha_{\mathsf{f}}/\kappa$ under the FA(2,1)-model. That is, under the FA(2,1)- model the parameter representing the amplitude of a simulated forcing effect is not α_{f} itself, but the ratio $\alpha_{\mathsf{f}}/\kappa$.

In addition, (2.3.16) shows the equivalence between $\sigma^2_{\xi_{\mathsf{f}}}$ and κ^2. Hence, testing the hypothesis $H_0 : \kappa = 0$ under the FA(2,1)-model is equivalent to testing the hypothesis $H_0 : \sigma_{\xi_{\mathsf{f}}} = 0$ under the ME model. It leads to the question as to whether tests for individual parameters of the factor model are available. To address this question let us investigate in more detail how the estimation procedure for a q-indicator p-factor model is carried out.

There is a vast range of literature devoted to factor analysis. Obviously, it is not possible within the confines of this work to give more than a cursory introduction to factor analysis. Readers who are interested in learning this topic in greater depth are referred to the sources given in the following discussion.

To begin with, factor analysis encompasses two major techniques of analyzing data: *exploratory factor analysis* (EFA) and *confirmatory factor analysis* (CFA). In EFA the structure of the factor model is not known or specified a priori. Data are used to help reveal the structure of the factor model. Therefore, EFA imposes no substantive constraints on the data; there are no restrictions on the pattern of relationships between observed and latent variables. Each common factor is assumed to affect every observed variable and that the common factors are either all correlated or uncorrelated. In CFA, on the other hand, the investigator has certain hy-

[6] In the terminology of factor analysis, parameters that are to be estimated are referred to as *free parameters*.

potheses about which factors are to be involved and which restrictions on the parameter space it implies. Depending on hypotheses the investigator has, values of some model parameters, e.g. factor loadings or variances of specific factors, can be specified in advance[7]. Formulating factor models with a certain structure gives rise to the question: How well do the empirical data conform to the hypothesized factor model? In other words, CFA can be viewed as a technique for theory testing. In this work, we will focus on the confirmatory factor analysis.

Depending on the number of the unique equations in the variance-covariance matrix of indicators and the number of free parameters, a factor model can be classified as underidentified, just-identified or overidentified.

If the number of the unique equations that is $q(q+1)/2$ is smaller than the number of free parameters, then a model is underidentified. By imposing constraints on some parameters, the number of free parameters can be reduced to $q(q+1)/2$. Such a model is called just-identified. Both previously discussed ME- and FA(2,1)-models are just-identified. An overidentified factor model is a model where the number of the unique equations is larger than the number of free parameters. At this point, a word of caution is needed: even if the number of free parameters is less than $q(q+1)/2$, this may not be sufficient in a specific case to make the model identified. Fixing some parameters to certain values or establishing equality constrains between parameters, researchers should check whether different sets of free parameter values, given the set of known (and constrained-equal) parameters, do not lead to the same hypothetical covariance matrix for the observed variables (see Definition 1). If it is a case, the model is not identified, and therefore it should be respecified.

If a model is identified (either just-identified or overidentified), estimates of model parameters are obtained by minimizing a discrepancy between the sample estimate of the unrestricted variance-covariance matrix $\boldsymbol{S}$ and the model's reproduced variance-covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector $\boldsymbol{\theta} = (\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$, with a subvector $\bar{\boldsymbol{\theta}}$ containing free parameters and $\boldsymbol{\theta}^*$ known parameters. The discrepancy is measured with a discrepancy function $F(\boldsymbol{\theta})$, conditional on the known parameters $\boldsymbol{\theta}^*$. There are a number of different discrepancy functions (Mulaik, 2010, Ch. 15). Under normality assumption of data, the following function is used:

$$F(\boldsymbol{\theta}) = \log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + \operatorname{tr}(\boldsymbol{S}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}) - \log|\boldsymbol{S}| - q, \tag{2.3.17}$$

[7]Note that parameters whose values are specified in advance in order to achieve the identifiability of a model are not a part of hypotheses.

To explain the idea behind this discrepancy function, we note first that the estimates $\widehat{\boldsymbol{\theta}}$, which minimize the discrepancy function F are the maximum-likelihood estimates, which maximize the logarithm of the likelihood function, conditional on the known parameters $\boldsymbol{\theta}^*$ (Jöreskog, 1969). Indeed, without a function of the observations the logarithm of the likelihood function under the null hypothesis $H_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ is given by

$$\log L(H_0) = -\frac{1}{2} \cdot (n-1) \cdot \Big\{ \log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + \mathrm{tr}(\boldsymbol{S}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}) \Big\}, \tag{2.3.18}$$

while under the alternative hypothesis H_A of unrestricted $\boldsymbol{\Sigma}$, i.e. $\boldsymbol{\Sigma} = \boldsymbol{S}$, the logarithm of the likelihood function is given by

$$\log L(H_A) = -\frac{1}{2} \cdot (n-1) \cdot \Big\{ \log|\boldsymbol{S}| + q \Big\}. \tag{2.3.19}$$

As we see, the F function is closely related to the log-likelihood ratio, used for testing the goodness of fit of the model's $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ to $\boldsymbol{S}$:

$$G = -2 \cdot (\log L(H_0) - \log L(H_A)) = (n-1) \cdot F(\widehat{\boldsymbol{\theta}}), \tag{2.3.20}$$

which is approximately distributed in large samples as chi-square with

$$df = q(q+1)/2 - m$$

degrees of freedom, where $q(q+1)/2$ is the number of the unique equations in the variance-covariance matrix of the indicators, and m is the number of distinct free parameters.

For just-identified models, the function $F(\widehat{\boldsymbol{\theta}})$, evaluated at the minimum, is equal to zero, since $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{S}$ and $\mathrm{tr}(\boldsymbol{S}\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})^{-1}) = q$. That is, a just-identified model has an exact solution in terms of the variances and covariances among indicators, but nothing is hypothesized and nothing can be tested.

For overidentified models, arising due to additional constraints imposed on some model parameters, at least one (free) parameter can be expressed by more than one distinct equation in terms of the variances and covariances of indicators. Therefore, the fit between $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})$ and the sample variance-covariance matrix, in general, will not be perfect, thus making it possible to assess the fit of the model to the data.

If the solution obtained is proper[8] and interpretable, the overall model

[8] One way to check whether a solution is proper or not is to look at the completely standardized solution. This type of solution standardizes the solution such that the variances of the latent factors *and* the indicators are one. Improper solutions are indicated by factor loading that do not lie between -1 and $+1$, and by specific variances that are greater than one (Sharma, 1996).

fit to the data can be assessed by means of the test statistic G, which is asymptotically χ^2 distributed with degrees of freedom equal to the difference between the number of the unique equations in $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and the number of free parameters. Note, failure to reject the null hypothesis is desired, as it leads to the conclusion that the hypothesized model with the resulting variance-covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ fits the data.

Finally, the estimation procedure also includes the estimation of the variances and covariances among the parameter estimates. According to the general theory, the ML estimates are consistent, jointly asymptotically normally distributed with the asymptotic variance expressed as being the inverse of the Fisher information. In confirmatory factor analysis, the information matrix is defined as follows (Bollen, 1989, p.135):

$$\frac{n-1}{2} \cdot E\left[\frac{\partial^2 F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]. \tag{2.3.21}$$

The inverse of (2.3.21), evaluated at the values for the parameters that minimize the F function, gives an estimate of the variance of the asymptotic distribution of the model estimates. Note that the inverse of the matrix of second-order derivatives of the $F(\boldsymbol{\theta})$ function is explicitly calculated if a Newton-Raphson algorithm is implemented whereas a quasi-Newton algorithm, e.g. the Fletcher-Powell algorithm implemented first by Jöreskog (1969), finds a close numerical approximation to it.

Provided that the information matrix is positive definite, each *estimated* parameter θ_i can be tested by means of the z statistic $\widehat{\theta}_i \Big/ \sqrt{\widehat{\mathrm{Var}}(\widehat{\theta}_i)}$, which has approximately a standard normal distribution. The results of tests that $\theta_i = 0$ are provided in form of two-sided p-values by all statistical packages designed to do confirmatory factor analysis, regardless of whether a model is just-identified or overidentified. This is comforting, since we are avoiding complex and burdensome algebraic calculations to derive an analytical expression for the estimated variance of $\widehat{\kappa} = s_{x_\mathrm{f} z} \Big/ \sqrt{s^2_{x_\mathrm{f}} - \sigma^2_{\delta_\mathrm{f}}}$. In addition, we also can construct the Wald confidence interval, defined in (2.3.12), for each parameter θ_i to test $H_0 : \theta_i = \theta_i^0$.

Regarding statistical software, factor analysis is mostly performed via commercial software, such as LISREL, Mplus, Amos. In this work, we employed the R package `sem`, `S`tructural `E`quation `M`odels, that is an open source alternative (see Fox, 2006; http://CRAN.R-project.org/package=sem). A distinguishing feature of the `sem` package is that it requires latent variable variances of 1 to be represented explicitly. This means that if a researcher wishes to estimate the ME model by means of the `sem` package instead

of using directly (2.3.6), he or she needs first to fit the FA(2,1)-model in (2.3.14), and then, take the ratio of the estimated factor loadings, $\widehat{\alpha}_{\mathsf{f}}/\widehat{\kappa}$. As known, the estimator of α_{f} and $\alpha_{\mathsf{f}}/\kappa$ under respective model is the same, namely $(s^2_{x_{\mathsf{f}}} - \sigma^2_{\delta_{\mathsf{f}}})/s_{x_{\mathsf{f}}z}$. But the availability of the estimated variances and covariances among *all* estimates of the FA(2,1)-model enables us to derive the confidence limits for $\alpha_{\mathsf{f}}/\kappa$ in another way than it was done under the ME model. The method, used in this work, is based on the Fieller method of finding the confidence interval of the ratio of two normal means (Franz, 2007; Cheng, 1999, Sec. 2.4.3).

The idea of the Fieller method is to use a **pivotal quantity**, that is a function of the data and parameters whose distribution does not depend on any unknown parameters. Here, for ease of exposition, we introduce a unified parameter, representing the amplitude of a forcing effect in a climate model, valid under both approaches. This is the ratio $\lambda_{11}/\lambda_{21}$, where λ_{11} is the loading of the first indicator x_{f} on the latent factor ξ_{f}, while λ_{21} is the loading of the second indicator z on the same latent factor. Under the ME model, i.e. under Approach 1, the ratio $\lambda_{11}/\lambda_{21}$ is equal to $\alpha_{\mathsf{f}}/1 = \alpha_{\mathsf{f}}$, while under the FA(2,1)- model, i.e. under Approach 2, it is equal to $\alpha_{\mathsf{f}}/\kappa$.

Recall that under assumption of normality of data, the estimates of loadings in a factor model are asymptotically jointly normally distributed with mean vector $\boldsymbol{\lambda}$ and the covariance matrix $\mathrm{Var}(\widehat{\boldsymbol{\lambda}})$. Thus for large samples we may assume that they are approximately normally distributed. Because the difference of (approximately) normal variables is also (approximately) normally distributed, the statistic $\widehat{\lambda}_{11} - (\lambda_{11}/\lambda_{21}) \cdot \widehat{\lambda}_{21}$ is approximately normally distributed with mean value zero and the variance $\sigma^2_{\widehat{\lambda}_{11}} + (\lambda_{11}/\lambda_{21})^2 \cdot \sigma^2_{\widehat{\lambda}_{21}} - 2 \cdot (\lambda_{11}/\lambda_{21}) \cdot \sigma_{\widehat{\lambda}_{11}\widehat{\lambda}_{21}}$. Replacing the unknown covariances by their estimates, we obtain the statistic

$$T_{\lambda_{11}/\lambda_{21}} = \frac{\widehat{\lambda}_{11} - (\lambda_{11}/\lambda_{21}) \cdot \widehat{\lambda}_{21}}{\sqrt{\widehat{\sigma}^2_{\widehat{\lambda}_{11}} + (\lambda_{11}/\lambda_{21})^2 \cdot \widehat{\sigma}^2_{\widehat{\lambda}_{21}} - 2 \cdot (\lambda_{11}/\lambda_{21}) \cdot \widehat{\sigma}_{\widehat{\lambda}_{11}\widehat{\lambda}_{21}}}} \tag{2.3.22}$$

that follows approximately the standard normal distribution[9].

[9]Note that under the ME model, the loading λ_{21} is κ that is fixed to 1. It implies that $\sigma^2_{\widehat{\lambda}_{21}}$ and $\sigma_{\widehat{\lambda}_{11}\widehat{\lambda}_{21}}$ are zero, which simplifies (2.3.22) to the test statistic in (2.3.11),

$$T = \frac{\widehat{\lambda}_{11} - (\lambda_{11}/\lambda_{21}) \cdot 1}{\widehat{\sigma}_{\widehat{\lambda}_{11}}} = \frac{\widehat{\alpha}_{\mathsf{f}} - \alpha_{\mathsf{f}}}{\widehat{\sigma}_{\widehat{\alpha}_{\mathsf{f}}}},$$

used to construct the Wald confidence interval in (2.3.12).

Consequently,

$$T^2_{\lambda_{11}/\lambda_{21}} = \frac{\left(\widehat{\lambda}_{11} - (\lambda_{11}/\lambda_{21}) \cdot \widehat{\lambda}_{21}\right)^2}{\sigma^2_{\widehat{\lambda}_{11}} + (\lambda_{11}/\lambda_{21})^2 \cdot \sigma^2_{\widehat{\lambda}_{21}} - 2 \cdot (\lambda_{11}/\lambda_{21}) \cdot \sigma_{\widehat{\lambda}_{11}\widehat{\lambda}_{21}}}, \qquad (2.3.23)$$

is approximately χ^2 distributed with 1 degree of freedom. The set

$$\left\{ \lambda_{11}/\lambda_{21} \;\middle|\; \frac{\left(\widehat{\lambda}_{11} - (\lambda_{11}/\lambda_{21}) \cdot \widehat{\lambda}_{21}\right)^2}{\widehat{\sigma}^2_{\widehat{\lambda}_{11}} + (\lambda_{11}/\lambda_{21})^2 \cdot \widehat{\sigma}^2_{\widehat{\lambda}_{21}} - 2 \cdot (\lambda_{11}/\lambda_{21}) \cdot \widehat{\sigma}_{\widehat{\lambda}_{11}\widehat{\lambda}_{21}}} \leq c_p, \right\} \qquad (2.3.24)$$

where c_p satisfies $P(\chi^2(1) \leq c_p) = 1 - p$, gives a $1 - p$ confidence region for $(\lambda_{11}/\lambda_{21})$. Expression (2.3.24) is equivalent to the quadratic inequality

$$a \cdot (\lambda_{11}/\lambda_{21})^2 - 2 \cdot b \cdot (\lambda_{11}/\lambda_{21}) + c \leq 0, \qquad (2.3.25)$$

where $a = \widehat{\lambda}^2_{21} - c_p \cdot \widehat{\sigma}^2_{\lambda_{21}}$, $b = \widehat{\lambda}_{11} \cdot \widehat{\lambda}_{21} - c_p \cdot \widehat{\sigma}_{\lambda_{11}\lambda_{21}}$ and $c = \widehat{\lambda}^2_{11} - c_p \cdot \widehat{\sigma}^2_{\lambda_{11}}$. The roots of this quadratic are either (i) both real or (ii) both complex. In case (i), the roots are

$$r_1 = -\sqrt{(b/a)^2 - (c/a)} + (b/a),$$

$$r_2 = \quad \sqrt{(b/a)^2 - (c/a)} + (b/a).$$

If $a > 0$ or in other words when the hypothesis $\widehat{\lambda}_{21} = 0$ is rejected at significance level p, the $1 - p$ confidence region is the interval between the roots, i.e.

$$CI_{\lambda_{11}/\lambda_{21}} = (r_1, r_2)\,,$$

and if $a < 0$, or equivalently when the hypothesis $\widehat{\lambda}_{21} = 0$ is not rejected at significance level p, it is the complement of this interval, that is, a confidence region which includes all values outside the roots but excludes the values between the roots, ("unbounded/exclusive" region):

$$CR_{\lambda_{11}/\lambda_{21}} = \{(-\infty, r_1)\,, (r_2, +\infty)\}\,.$$

In case (ii), inequality (2.3.25) is satisfied only when $a < 0$ for all $\lambda_{11}/\lambda_{21}$, which means that the confidence interval is the whole real line ("unbounded" region):

$$CR_{\lambda_{11}/\lambda_{21}} = (-\infty, +\infty)\,.$$

As we see, the proposed method is able to generate unbounded regions, which means that the zero confidence level effect due to Gleser and Hwang's theorem does not persist. This fact, together with the possibility to test all individual parameters, makes use of the FA(2,1)-model more advantageous, compared to the ME model[10].

2.3.3 Relation to other studies

Detection and attribution studies

The objective of detection and attribution (D&A) studies, generally known as "optimal fingerprinting", is the assessment of the amplitude of the response to external climate forcings in temperature reconstructions. As in SUN12, these studies often assume that forcing effects (called *fingerprints of forcings*) are contaminated by noise.

Using the notation of the present work, a statistical model used in the D&A studies for assessing the amplitude of the *overall* forced response in temperature reconstructions can be expressed as follows:

$$\begin{cases} x_{\mathsf{total}\ t} &= \xi_{\mathsf{total}\ t} &+ \delta_{\mathsf{total}\ t}, \\ z_t &= \beta \cdot \xi_{\mathsf{total}\ t} &+ \underbrace{\gamma_t}_{=\eta_{\mathsf{internal}\ t}+\epsilon_t} \end{cases} \tag{2.3.26}$$

or equivalently (in a form more familiar in climate D&A studies)

$$z_t = \beta \cdot (x_{\mathsf{total}\ t} - \delta_{\mathsf{f}\ t}) + \gamma_t.$$

The model above is estimated using the *total least squares* (TLS) approach (see Allen and Stott, 2003), which corresponds to a ME model under the identifiability condition that the ratio of the noise variances, $\sigma_\gamma^2/\sigma_{\delta_\mathsf{f}}^2$, is known. While not obvious, the ME model in (2.3.26) has a direct bearing on the statistical framework of SUN12. To see it let us formulate an FA(2,1)-model for a climate model driven by *all* external forcings using the SUN12 specifications:

[10] Frankly speaking, it is possible to apply the Fieller method even under the ME model. Indeed, we could denote $(s_{x_\mathsf{f}}^2 - \sigma_{\delta_\mathsf{f}}^2)$ by λ_{11}, and $s_{x_\mathsf{f} z}$ as λ_{21}, and use the fact that the limiting distribution of $\sqrt{n}(s_{x_\mathsf{f}}^2 - \sigma_{x_\mathsf{f}}^2, s_{x_\mathsf{f} z} - \sigma_{x_\mathsf{f} z})^T$, where x_f and z are normally distributed random variables is a normal distribution with the properties defined according to Fuller, Appendix 1.C, Corollary 1.C. 1. However, constructing a confidence set for α_f in this way does not solve all problems associated with the ME model, in particular, it does not aid in testing the hypothesis $H_0 : \sigma_{\xi_\mathsf{f}}^2 = 0$. Therefore, it was decided not to apply this method under the ME model.

$$\begin{cases} x_{\mathsf{total}\ t} &= \alpha_{\mathsf{total}} \cdot \xi_{\mathsf{total}\ t} &+ & \delta_{\mathsf{total}\ t} \\ z_t &= \beta \cdot \xi_{\mathsf{total}\ t} &+ & \underbrace{\gamma_t}_{=\eta_{\mathsf{internal}\ t}+\epsilon_t} \end{cases} . \tag{2.3.27}$$

The identifiability of the above model is achieved by letting the factor loading for one of the indicators to be 1, or by letting the variance of $\xi_{\mathsf{total}\ t}$ to be 1. As we see, model (2.3.26) was derived from model (2.3.27) by setting $\alpha_{\mathsf{total}} = 1$. Having estimated β and the asymptotic variance of $\widehat{\beta}$, one can construct the $1-p$ Wald confidence interval for β according to (2.3.12). However, if the correlation between x_{total} and z is low, arbitrarily large deviations from the intended confidence level are possible. In this situation, the use the FA(2,1)-model in (2.3.27) instead of model (2.3.26) can be recommended. Letting $\mathrm{Var}(\xi_{\mathsf{total}}) = 1$ in (2.3.27), the amplitude of the overall forcing effect in z is then represented by the ratio $\beta/\alpha_{\mathsf{total}}$. Consequently, the associated confidence region can be constructed according to the Fieller method (see (2.3.25)), which does not suffer from the zero confidence level effect.

It can also be recommended to replace the identifiability condition of known variance ratio by the more realistic condition of known $\sigma^2_{\delta_{\mathsf{f}}}$. Since replacing one condition by another leads necessarily to a new estimator, it can be an appropriate topic for further research to compare the performance of two estimators, applied to the same climate data.

We conclude by remarking that the FA(2,1)-model can be used even for estimating the amplitude of the forced response to a particular forcing f in temperature reconstructions. In that case, the FA(2,1)-model in (2.3.14), formulated for a climate model driven by a particular forcing, should be applied. Under this model, isolating the true forcing effect associated with a particular forcing, ξ_{f}, from the true total forcing effect makes a correct value of the amplitude of the forced response in observations known. It equals 1. The parameter representing this amplitude is the ratio $\kappa/\alpha_{\mathsf{f}}$, i.e. the inverse of $\alpha_{\mathsf{f}}/\kappa$. Note that if contributions of several *individual* forcings are in focus, a statistical model employed in D&A studies is a ME model with vector explanatory variables when the entire error covariance structure is known up to a scalar multiple[11] (Allen and Stott, 2003; see also Fuller,

[11] If fingerprints are assumed to be noise free then model (2.3.28) reduces to an ordinary multiple regression model.

1987, Sec. 2.3):

$$z_t = \sum_{\mathtt{f}_i=1}^{m} \beta_i \cdot (x_{\mathtt{f}_i t} - \delta_{\mathtt{f}_i t}) + \gamma_t. \tag{2.3.28}$$

As follows from (2.3.28), one of the model assumptions is that interactions between forcings, influencing the climate simultaneously, are negligible, which explains the absence of possible interaction terms between the fingerprints of individual forcings, $(x_{\mathtt{f}_i t} - \delta_{\mathtt{f}_i t})$, in the model. Under the FA(2,1)-model in (2.3.14) this assumption is relaxed.

Relation to the analysis by Tingley et al. (2015)
Assuming that the true temperature is observable, the statistical model formulated by Tingley et al. (2015) is the following:

$$\begin{cases} Y_t &= F_t^M + U_t^M, \\ C_t &= F_t^P + U_t^P \\ F_t^M &= \beta_1 \cdot F_t^P + D_t \end{cases} \tag{2.3.29}$$

or equivalently

$$\begin{cases} Y_t &= \beta_1 \cdot F_t^P + D_t + U_t^M \\ C_t &= F_t^P + U_t^P \end{cases}$$

where, using the notations of the present work,

$$Y_t \equiv x_{\mathtt{f}\,t}, \quad C_t \equiv \tau_t, \quad F_t^P \equiv \xi_{\mathsf{total}\,t}, \quad F_t^M \equiv \alpha_{\mathtt{f}} \cdot \xi_{\mathtt{f}\,t},$$

$$U_t^M \equiv \delta_{\mathtt{f}\,t}, \quad U_t^P \equiv \eta_{\mathsf{internal}\,t},$$

and the variable D_t, independent from F_t^P, represents the disrepancy between the two forced series, and is assumed to be normally distributed with zero mean and variance σ_D^2. We omit the description of the distributional properties of the remaining variables.

The model in (2.3.29) is known as a ME model with an error in the equation (see Cheng, 1999, Sec. 1.5). For the standard ME model, defined as in (2.3.1), σ_D^2 is 0, which, however, is not the only difference between these models. As we see, the decomposition of the true temperature, C_t, does not coincide with the decomposition of τ_t in SUN12, where the effect of a particular forcing `f` is isolated from the total forced response:

$$\tau_t = \xi_{\mathtt{f}\,t} + \underbrace{(\xi_{\mathsf{total}\perp\mathtt{f}\,t} + \eta_{\mathsf{internal}\,t})}_{=\eta_t}.$$

The consequence of modeling a simulated forcing effect as a function of the true *total* forcing effect, as in (2.3.29), is that for a climate model driven by some (not all) external forcings a correct value of β_1 is not known, implying that the value of unity for the slope β_1 is not necessarily a correct value. Indeed, these forcings might be weak to their nature and thus not capable to generate substantial temperature changes with the same (high) amplitude as F^P has. Therefore, observing $\widehat{\beta}_1 \approx 1$ in this situation would rather indicate a discrepancy than an agreement between the reconstructions of the involved individual forcings and their real-world counterparts. Instead, it would be of little surprise to observe quite small values of $\widehat{\beta}_1$ for a climate model driven by weak forcings. But this in no way means that the amplitude of the simulated forcing effect, F^M, in a climate model is too small.

Finally, we note that inclusion of the discrepancy term D in the ME model in (2.3.1), or equivalently in the FA(2,1)-model in (2.3.14), will lead to unidentifiable models. For Bayesian approach, the issue of identifiability does not arise, making it possible to estimate more parameters than can be identified from a sample variance-covariance matrix. Therefore, it is of interest to compare the performance of model (2.3.29) with C_t replaced by observations and the ME model in (2.3.1), or equivalently the FA(2,1)-model in (2.3.14), by applying them to the same climate data including simulations generated by climate models driven by all external forcings.

2.4 Models with one latent factor. Heteroscedasticity. Approach 1: $\kappa = 1$ and Approach 2: $\mathbf{Var}(\xi_{\mathsf{f}}) = 1$

In the preceding discussion it was assumed that the proxy variance is constant over time. Now, let the proxy series z_t have a time-varying precision, $\sigma_\epsilon^2(t)$. It implies that the specific factor ν in the ME- and FA(2,1)-models has the following time-varying variance:

$$\sigma_\nu^2(t) = \sigma_\eta^2 + \sigma_\epsilon^2(t) = \sigma^2_{\xi_{\mathsf{total}\perp\mathsf{f}}} + \sigma^2_{\eta_{\mathsf{internal}}} + \sigma_\epsilon^2(t).$$

To take the heteroscedasticity in the proxy series $\{z_t\}$ into account, we suggest to replace $s_{x_{\mathsf{f}}z}$ and s_z^2 by weighted versions, given by:

$$\begin{aligned} s^{(w)}_{x_{\mathsf{f}}\,z} &= \frac{\sum_{t=1}^n w_t \cdot (x_{\mathsf{f}\,t} - \mu_{x_{\mathsf{f}}})(z_t - \mu_z)}{\sum_{t=1}^n w_t} \\ s_z^{2(\breve{w})} &= \frac{\sum_{t=1}^n \breve{w}_t \cdot (z_t - \mu_z)^2}{\sum_{t=1}^n \breve{w}_t}. \end{aligned} \tag{2.4.1}$$

where the mean values μ_{x_f} and μ_z are assumed to be known, $\{w_t\}$ and $\{\breve{w}_t\}$ are suitable sets of weights (not necessarily summing to 1). Note that if $w_i = w_j$ and $\breve{w}_i = \breve{w}_j$ for all i, j such that $i \neq j$, the weighted statistical functions above become ordinary maximum likelihood estimates of the population variance and covariance, i.e.

$$\begin{aligned} s_{x_\mathsf{f}\, z} &= \frac{\sum_{t=1}^n (x_{\mathsf{f}\, t} - \mu_{x_\mathsf{f}})(z_t - \mu_z)}{n} \\ s_z^2 &= \frac{\sum_{t=1}^n (z_t - \mu_z)^2}{n}. \end{aligned}$$

In practice, the mean values μ_{x_f} and μ_z, are not known, and therefore it is natural to replace them by the weighted averages, $\bar{x}_\mathsf{f}^{(w)} = \sum_{t=1}^n w_t x_{\mathsf{f}t} \Big/ \sum_{t=1}^n w_t$, etc. It can be shown that just as with the ordinary statistical functions, where μ_{x_f} and μ_z are replaced by the ordinary averages, the use of the estimated mean values introduces a bias, which however can be corrected by means of the correction factors

$$W = 1 - \frac{\sum_{t=1}^n w_t^2}{\left(\sum_{t=1}^n w_t\right)^2}, \quad \text{and} \quad \breve{W} = 1 - \frac{\sum_{t=1}^n \breve{w}_t^2}{\left(\sum_{t=1}^n \breve{w}_t\right)^2},$$

respectively. If all w_t respectively $\breve{w}_t$ are equal, the corresponding correction factor above simplifies to the standard correction factor $1 - 1/n$, motivated in ordinary sample covariances and ordinary sample variances. Since both W and $\breve{W}$ tend to 1 as $n \to \infty$ (provided that $\max\limits_{1 \le t \le n} \left(w_t / \sum_{t=1}^n w_t\right)$ respective $\max\limits_{1 \le t \le n} \left(\breve{w}_t / \sum_{t=1}^n \breve{w}_t\right)$ goes to 0 as $n \to \infty$), the bias in both weighted and ordinary estimates will be negligible when the sample size, n, is large. Referring to the fact that in the present analysis we are dealing with large samples, the correction factors W and $\breve{W}$ will be omitted from calculations.

To see what parameters the weighted functions in (2.4.1) are expected to estimate, we calculate their expectations:

$$E\left[s_{x_\mathsf{f}\, z}^{(w)}\right] = E\left[\frac{\sum_{t=1}^n w_t \left(x_{\mathsf{f}\, t} - \mu_\mathsf{f}\right)\left(z_t - \mu_z\right)}{\sum_{t=1}^n w_t}\right] =$$

$$= \sigma_{x_\mathsf{f} z} = \begin{cases} \alpha_\mathsf{f} \cdot \sigma_{\xi_\mathsf{f}}^2 & \text{under the ME model} \\ & \text{with} \quad \kappa = 1 \\ \\ \alpha_\mathsf{f} \cdot \kappa & \text{under the FA(2,1)- model} \\ & \text{with } \sigma_{\xi_\mathsf{f}}^2 = 1, \end{cases} \tag{2.4.2}$$

and

$$E\left[s_z^{2(\check{w})}\right] = E\left[\frac{\sum_{t=1}^{n}\check{w}_t\left(z_t-\mu_z\right)^2}{\sum_{t=1}^{n}\check{w}_t}\right] = E\left[\frac{\sum_{t=1}^{n}\check{w}_t\left(\kappa\cdot\xi_{\mathsf{f}\,t}+\nu_t\right)^2}{\sum_{t=1}^{n}\check{w}_t}\right] =$$

$$= \begin{cases} \sigma_{\xi_{\mathsf{f}}}^2 + \sigma_\nu^{2(\check{w})} & \text{under the ME model with } \kappa = 1 \\ \\ \kappa^2 + \sigma_\nu^{2(\check{w})} & \text{under the FA(2,1)- model with} \sigma_{\xi_{\mathsf{f}}}^2 = 1, \end{cases} \tag{2.4.3}$$

where $\sigma_\nu^{2(\check{w})} = \frac{\sum_t \check{w}_t\cdot\sigma_\nu^2(t)}{\sum_t \check{w}_t}$, that is, we estimate the weighted average variability of z over the whole time period.

Using the weighted functions the consistent estimates of the model parameters become:

$$\begin{array}{lcll} \underline{\text{ME}} & & \underline{\text{FA(2,1)}} & \\ \widehat{\alpha}_{\mathsf{f}} & \leftrightarrow & \widehat{\alpha}_{\mathsf{f}}/\widehat{\kappa} & = \dfrac{s_{x_{\mathsf{f}}}^2 - \sigma_{\delta_{\mathsf{f}}}^2}{s_{x_{\mathsf{f}}z}^{(w)}}, \\ \\ \widehat{\sigma}_{\xi_{\mathsf{f}}}^2 & \leftrightarrow & \widehat{\kappa}^2 & = \left(s_{x_{\mathsf{f}}z}^{(w)}\right)^2 \Big/ \left(s_{x_{\mathsf{f}}}^2 - \sigma_{\delta_{\mathsf{f}}}^2\right), \\ \\ \widehat{\sigma}_\nu^{2(\check{w})} & = & \widehat{\sigma}_\nu^{2(\check{w})} & = s_z^{2(\check{w})} - \left(s_{x_{\mathsf{f}}z}^{(w)}\right)^2 \Big/ \left(s_{x_{\mathsf{f}}}^2 - \sigma_{\delta_{\mathsf{f}}}^2\right), \end{array} \tag{2.4.4}$$

provided the following side conditions are met: (1) $s_{x_{\mathsf{f}}\,z}^{(w)} \neq 0$, (2) $s_{x_{\mathsf{f}}}^2 > \sigma_{\delta_{\mathsf{f}}}^2$ and (3) $s_z^{2(\check{w})} - \left(s_{x_{\mathsf{f}}z}^{(w)}\right)^2 \Big/ \left(s_{x_{\mathsf{f}}}^2 - \sigma_{\delta_{\mathsf{f}}}^2\right) \geq 0$.

As follows from (2.4.2), the expected value of the weighted covariance is the same as the expected value of its ordinary counterpart. Hence, we could use the conventional estimator that is $(s_{x_{\mathsf{f}}}^2 - \sigma_{\delta_{\mathsf{f}}}^2)/s_{x_{\mathsf{f}}z}$ because it remains consistent even in the presence of heteroscedasticity. Nevetheless, its precision will be lower, especially when the time-varying noise in the proxy dominates the time-constant variability of the forcing-related component ξ_{f}, i.e. when $\sigma_{\xi_{\mathsf{f}}}^2 \approx 0$ or equivalently $\kappa \approx 0$, depending on identification approach. By choosing the weights for $s_{x_{\mathsf{f}}z}^{(w)}$ such that its variance is minimized under the particular circumstances, we obtain the most efficient estimator for the population covariance, and thereby the most efficient estimator for $\lambda_{11}/\lambda_{21}$, representing the amplitude of the forcing effect in a climate model.

Noting that under the ME model $\mathrm{Var}\left(s^{(w)}_{x_{\mathsf{f}}z}\right)$ is given by:

$$\mathrm{Var}\left(\sum_{t=1}^{n} w'_t\,(x_{\mathsf{f}t}-\mu_{x_{\mathsf{f}}})\,(z_t-\mu_z)\right)=\mathrm{Var}\left(\sum_{t=1}^{n} w'_t(\alpha_{\mathsf{f}}\cdot\xi_{\mathsf{f}t}+\delta_{\mathsf{f}t})(\xi_{\mathsf{f}t}+\nu_t)\right)= \tag{2.4.5}$$

$$=\underbrace{\alpha^2_{\mathsf{f}}\cdot 2\cdot(\sigma^2_{\xi_{\mathsf{f}}})^2\cdot\sum_{t=1}^{n} w'^2_t+\alpha^2_{\mathsf{f}}\cdot\sigma^2_{\xi_{\mathsf{f}}}\cdot\sum_{t=1}^{n} w'^2_t\cdot\sigma^2_\nu(t)+\sigma^2_{\delta_{\mathsf{f}}}\cdot\sigma^2_{\xi_{\mathsf{f}}}\cdot\sum_{t=1}^{n} w'^2_t}_{\text{negligible when }\sigma^2_{\xi_{\mathsf{f}}}\approx 0}+$$

$$+\sigma^2_{\delta_{\mathsf{f}}}\cdot\sum_{t=1}^{n} w'^2_t\cdot\sigma^2_\nu(t)\approx$$

$$\approx\sigma^2_{\delta_{\mathsf{f}}}\cdot\sum_{t=1}^{n} w'^2_t\cdot\sigma^2_\nu(t)\,,$$

where $\sigma^2_{\xi_{\mathsf{f}}}$ is neglected, $w'_t=w_t/\sum_t w_t$ are the normalized weights, summing to 1, the problem of determining the weights essentially reduces to

$$\min_{w'_t}\quad\sum_{t=1}^{n} w'^2_t\cdot\sigma^2_\nu(t), \tag{2.4.6}$$

$$\text{subject to}\quad\sum_{t=1}^{n} w'_t=1.$$

Solving for w'_t, we obtain the weights proportional to $1/\sigma^2_\nu(t)$, i.e.

$$w'_t=\frac{1/\sigma^2_\nu(t)}{\sum_{t=1}^{n}1/\sigma^2_\nu(t)}, \tag{2.4.7}$$

valid under both ME and FA(2,1)-models, i.e. under both approaches. Hence, the original weights, w_t, are

$$w_t=\frac{1}{\sigma^2_\nu(t)}=\frac{1}{\sigma^2_\eta+\sigma^2_\epsilon(t)}. \tag{2.4.8}$$

In analogy with the U_R statistic, we may constrain the weights to be ≤ 1 by introducing the constant factor σ^2_η:

$$w_t=\frac{\sigma^2_\eta}{\sigma^2_\eta+\sigma^2_\epsilon(t)} \tag{2.4.9}$$

with equality when $\sigma^2_\epsilon(t)=0$. In this context, it should be pointed out that the weights above and the weights in U_R, denoted $\widetilde{w}_t$ in SUN12, differ, the

latter being the square root of the former. The explanation is that they were motivated by different principles, for different purposes. In case of the U_R-statistic, the main principle was to scale-transform $(z_t - \mu_z)$ by the weights such that $w_t(z_t - \mu_z)$ got the same variance for all time units t.

Regarding the weights for the weighted variance of z, $s_z^{2(\breve{w})}$, they were chosen in the simular way as the weights for $s_{x_\mathsf{f} z}^{(w)}$. Calculations resulted in

$$\breve{w}_t = w_t^2 = \widetilde{w}_t^4 = \left(\frac{\sigma_\eta^2}{\sigma_\nu^2(t)} \right)^2 . \tag{2.4.10}$$

Under the ME model specification, one can construct a large sample confidence interval according to (2.3.12). To this end, the following estimate of the variance of the limiting distribution of $\widehat{\alpha}_\mathsf{f}$ can be used:

$$\widehat{\mathrm{Var}}(\widehat{\alpha}_\mathsf{f}) = \frac{1}{n \cdot \left(s_{x_\mathsf{f} z}^{(w)} \right)^2} \cdot \left(s_{x_\mathsf{f}}^2 \cdot \left(\sigma_{\delta_\mathsf{f}}^2 + \widehat{\alpha}_\mathsf{f} \cdot \widehat{\sigma}_\nu^{2(\breve{w})} \right) + \sigma_{\delta_\mathsf{f}}^4 \right). \tag{2.4.11}$$

Under homoscedasticity (2.4.11) reduces to $\widehat{\mathrm{Var}}(\widehat{\alpha}_\mathsf{f})$ in (2.3.10).

Another important aspect to discuss is the fact that the weights should preferably be determined and estimated prior to analyzing the data. This requirement makes us recall that the variance of η includes two other variances, namely $\sigma^2_{\xi_{\mathsf{total}\perp\mathsf{f}}}$ and $\sigma^2_{\eta_{\mathsf{internal}}}$, where the former variance is not estimable (see the discussion in the beginning of Sec. 2.3.1). A possible remedy is to replace σ_η^2 by $\sigma^2_{\eta_{\mathsf{internal}}}$, which leads to the following weights:

$$w_t = \frac{\sigma^2_{\eta_{\mathsf{internal}}}}{\sigma^2_{\eta_{\mathsf{internal}}} + \sigma_\epsilon^2(t)} \tag{2.4.12}$$

At this point, the theoretical discussion regarding possible estimators of the amplitude of a forcing effect in a climate model could be finished, but the suggestion to replace σ_η^2 by $\sigma^2_{\eta_{\mathsf{internal}}}$ gives us an idea about how the ME and FA(2,1) models can be modified such that the weights in (2.4.12) will be motivated. Bearing in mind the way of decomposing the total forcing effect described in Eq. (2.2.4), this aim can be achieved by moving the orthogonal complement $\xi_{\mathsf{total}\perp\mathsf{f}}$ from the noise term η and by considering it as a second latent factor instead.

Combining Eq. (2.2.4) with the expression for the *mean-centered* simulated temperature leads to the following equation system:

$$\begin{cases} x_{\mathsf{f}\,t} = & \alpha_\mathsf{f} \cdot \xi_{\mathsf{f}\,t} & + & 0 \cdot \xi_{\mathsf{total}\perp\mathsf{f}\,t} & + & \delta_{\mathsf{f}\,t} \\ \tau_t = & \kappa \cdot \xi_{\mathsf{f}\,t} & + & \xi_{\mathsf{total}\perp\mathsf{f}\,t} & + & \eta_{\mathsf{internal}\,t}. \end{cases} \tag{2.4.13}$$

This is a two-indicator two-factor model, abbr. FA(2,2), where the matrix of the factor loadings has a specified pattern. More precisely, the loading for the first indicator on the second latent factor, λ_{12}, is fixed to zero, while the loading for the second indicator on the second latent factor, λ_{22}, is fixed to 1. This particular pattern can be explained by (1) our conviction that the first indicator depends only on the first latent factor and (2) by our way of decomposing the total forcing effect. This model is an excellent example of purely confirmatory factor analysis; the underlying theory determines which factors are to be involved and which restrictions on the parameter space it implies. Our aim is to find solutions which conform to this pattern.

Let us discuss new models with two latent factors under each approach. For the sake of simplicity, we start with assuming homoscedasticity.

2.5 Models with two latent factors. Homoscedasticity

2.5.1 Approach 1: $\kappa = 1$

By replacing τ in the FA(2,2)-model in (2.4.13) by the proxy z with a time-constant precision and applying the first identification approach, we obtain:

$$\begin{cases} x_{\mathsf{f}\,t} = & \alpha_{\mathsf{f}} \cdot \xi_{\mathsf{f}\,t} & + & 0 \cdot \xi_{\mathsf{total}\perp\mathsf{f}\,t} & + & \delta_{\mathsf{f}\,t} \\ z_t = & & \xi_{\mathsf{f}\,t} & + & \xi_{\mathsf{total}\perp\mathsf{f}\,t} & + & \underbrace{\gamma_t}_{=\eta_{\mathsf{internal}}+\epsilon_t} \end{cases}, \tag{2.5.1}$$

where $\delta_{\mathsf{f}\,t}$ and γ_t are assumed to be normally distributed and mutually uncorrelated random variables with zero mean and variances σ_δ^2 and $\sigma_\gamma^2 = \sigma_{\eta_{\mathsf{internal}}}^2 + \sigma_\epsilon^2$, respectively. The latent factors, ξ_{f} and $\xi_{\mathsf{total}\perp\mathsf{f}}$, can be assumed to be either fixed or random. Specifically:

`Structural case:` $(\xi_{\mathsf{f}\,t}, \xi_{\mathsf{total}\perp\mathsf{f}\,t})^T \sim \mathrm{N}\left(\mathbf{0}, \mathrm{diag}(\sigma_{\xi_{\mathsf{f}}}^2, \sigma_{\xi_{\mathsf{total}\perp\mathsf{f}}}^2\right),$

`Functional case:` $\frac{1}{n}\sum_t \xi_{\mathsf{f}\,t} = 0, \quad \sigma_{\xi_{\mathsf{f}}}^2 \equiv \frac{1}{n}\sum_t \xi_{\mathsf{f}\,t}^2 > 0\ ;$

$\frac{1}{n}\sum_t \xi_{\xi_{\mathsf{total}\perp\mathsf{f}}\,t} = 0, \quad \sigma_{\xi_{\mathsf{total}\perp\mathsf{f}}}^2 \equiv \frac{1}{n}\sum_t \xi_{\xi_{\mathsf{total}\perp\mathsf{f}}\,t}^2 > 0.$

In order to be able to construct a confidence set for α_{f} according to the Fieller method, we reparameterize the model in the following way:

1. Define standardized latent factors: $\xi'_{\mathsf{f}} = \xi_{\mathsf{f}}/\sigma_{\xi_{\mathsf{f}}}$ and $\xi'_{\mathsf{total}\perp\mathsf{f}} = \xi_{\mathsf{total}\perp\mathsf{f}}/\sigma_{\mathsf{total}\perp\mathsf{f}}$.
2. Insert $\sigma_{\xi_{\mathsf{f}}} \cdot \xi'_{\mathsf{f}}$ and $\sigma_{\mathsf{total}\perp\mathsf{f}} \cdot \xi'_{\mathsf{total}\perp\mathsf{f}}$ instead of ξ_{f} and $\xi_{\mathsf{total}\perp\mathsf{f}}$, respectively, in model (2.5.1).

That is, we may rewrite model (2.5.1) as follows[12]:

$$\begin{cases} x_{\mathsf{f}\,t} = & \lambda_{11} \cdot \xi'_{\mathsf{f}\,t} + & 0 \cdot \xi'_{\mathsf{total}\perp\mathsf{f}} & + & \delta_{\mathsf{f}\,t} \\ z_t = & \lambda_{21} \cdot \xi'_{\mathsf{f}\,t} + \lambda_{22} \cdot \xi'_{\mathsf{total}\perp\mathsf{f}} & + & \gamma_t, \end{cases} \tag{2.5.2}$$

where the factor loadings are

$$\lambda_{11} = \alpha_{\mathsf{f}} \cdot \sigma_{\xi_{\mathsf{f}}}, \quad \lambda_{21} = \sigma_{\xi_{\mathsf{f}}}, \quad \text{and } \lambda_{22} = \sigma_{\xi_{\mathsf{total}\perp\mathsf{f}}}.$$

As we see, the ratio $\lambda_{11}/\lambda_{21}$ gives us back the parameter α_{f}, though λ_{21} is not equal to 1 as under model (2.5.1). To determine the identifiability condition, consider the model's reproduced variance-covariance matrix of the indicators:

$$\begin{aligned} \sigma^2_{x_{\mathsf{f}}} &= \lambda^2_{11} + \sigma^2_{\delta_{\mathsf{f}}} \\ \sigma_{x_{\mathsf{f}},z} &= \lambda_{11} \cdot \lambda_{21} \\ \sigma^2_z &= \lambda^2_{21} + \lambda^2_{22} + \sigma^2_\gamma. \end{aligned} \tag{2.5.3}$$

Since there are three unique equations, we conclude that only three functions of the five parameters can be estimated. To obtain $\widehat{\alpha}_{\mathsf{f}}$ we need to estimate λ_{11} and λ_{21}. The loading λ_{22} should also be estimated, since we do not have any a priori knowledge about its value. Consequently, to permit the identifiability of the model, both specific variances, $\sigma^2_{\delta_{\mathsf{f}}}$ and σ^2_γ, should be treated as known parameters. As concluded earlier, it is realistic to determine their values from the sources independent from the variance-covariance matrix of the indicators.

Treating the specific variances as known, the standardized FA(2,2)-model in (2.5.2) becomes just-identified, abbr. `j.i.FA(2,2)`. As follows from (2.5.3), the estimator of $\lambda_{11}/\lambda_{21}$ is the same as under the FA(2,1)-model (as well as under the ME model), provided that three side conditions are fulfilled: `(1)` $s_{x_{\mathsf{f}}z} \neq 0$, `(2)` $s^2_{x_{\mathsf{f}}} - \sigma^2_{\delta_{\mathsf{f}}} > 0$ and `(3)` $s^2_z - (s_{x_{\mathsf{f}}z})^2/(s^2_{x_{\mathsf{f}}} - \sigma^2_{\delta_{\mathsf{f}}}) - \sigma^2_\gamma \geq 0$. The latter condition is based on the requirement that $\widehat{\lambda}^2_{22}$, representing the variance of the second latent factor, must be nonnegative.

Hence, the `j.i.FA(2,2)`-model does not provide a new estimator of the amplitude of a forcing effect. However, if the third side condition is not satisfied, the model can be simplified, which naturally leads to a new estimator of $\lambda_{11}/\lambda_{21}$. The rejection of the third condition is interpreted as the variance of $\xi_{\mathsf{total}\perp\mathsf{f}}$ is small, and estimated as zero. Due to the absence of

[12] Henceforth we refer to factor models with standardized latent factors as standardized factor models.

complete knowledge about the climatological properties of $\xi_{\mathsf{total}\perp\mathsf{f}}$, the simplification is motivated even from the climatological point of view. Imposing the restriction $\lambda_{22} = 0$ leads to a one-factor model, FA(2,1):

$$\begin{cases} x_{\mathsf{f}\,t} = & \underbrace{\lambda_{11}}_{=\alpha_{\mathsf{f}}\cdot\sigma_{\xi_{\mathsf{f}}}} \cdot \xi'_{\mathsf{f}\,t} + & \delta_{\mathsf{f}\,t} \\ z_t = & \underbrace{\lambda_{21}}_{=\sigma_{\xi_{\mathsf{f}}}} \cdot \xi'_{\mathsf{f}\,t} + & \gamma_t, \end{cases} \tag{2.5.4}$$

where the specific variances are still considered as known, i.e. the model parameters are λ_{11} and λ_{21}. Since the number of the model parameters is less than the number of the unique equations in the model's variance-covariance matrix that is three (set $\lambda_{22} = 0$ in (2.5.3) to see it), the model is overidentified, abbr. `o.i.FA(2,1)`. The estimates of λ_{11} and λ_{21} are obtained by minimizing the discrepancy function given in (2.3.17) conditional on the known model parameter, i.e. $\lambda_{22} = 0$. If an admissible solution is obtained, the overall model fit can be assessed statistically by means of the G statistic, defined in (2.3.20). Naturally, the R package `sem` provides the value of the G statistic with the associated p-value.

Unfortunately, the usefulness of the statistic G has frequently been questioned by many researchres because in large samples even small differences between $\boldsymbol{S}$ and $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})$ will be statistically significant although the differences may not be practically meaningful. This is because the larger the sample size, the better approximation of the statistic's distribution to the chi-squared distribution is, implying the higher power of the statistic against any slight difference between $\boldsymbol{S}$ and $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})$. Consequently, a number of goodness-of-fit indices, serving as heuristic measures of model fit, have been proposed in the factor analysis literature (see for example Mulaik, 2010, Sec. 15.3.15; Sharma, 1996, sec. 6.4.3). Some fit indices are designed to provide a summary measure of the residual matrix, which is the difference between the sample and the estimated covariance matrix, i.e. $\boldsymbol{S} - \boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})$. Such indices are called *absolut* fit indices: no reference model is used to assess the amount of increment in model fit (Hu&Bentler, 1998). Here, we will use: a goodness-of-fit index (*GFI*); *GFI* adjusted for degrees of freedom (*AGFI*), and *standardized root- mean-square residual* (*SRMR*). They are defined as follows :

$$GFI = 1 - \frac{\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S} - \boldsymbol{I})^2}{\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S})^2}, \tag{2.5.5}$$

$$AGFI = 1 - \frac{q(q+1)}{2 \cdot df}\,(1 - GFI)\,, \tag{2.5.6}$$

where df are the degrees of freedom, q is the number of indicators, and finally

$$SRMR = \sqrt{\frac{\sum_{i=1}^{q}\sum_{j=1}^{i}\left[(s_{ij} - \widehat{\sigma}_{ij}) \Big/ (s_{ii}s_{jj})\right]^2}{q(q+1)/2}}, \tag{2.5.7}$$

where s_{ij} := observed variances, $\widehat{\sigma}_{ij}$:= reproduced covariances, s_{ii} and s_{jj} := observed standard deviations.

Regarding the cutoff values of the indices, the following rules of thumb are recommended. The GFI for good-fitting models should be greater than 0.90, while for the $AGFI$ the suggested cutoff value is 0.8 (Sharma, 1996). However, it has been shown that the expected value of GFI, and consequently $AGFI$, tends to increase with sample size.

In case with the $SRMR$, perfect model fit is indicated by $SRMR = 0$. Consequently, the larger the $SRMR$, the less fit between the model and the data. According to Hu and Bentler (1999), a cutoff value close to 0.08 for $SRMR$ indicates a good fit. Moreover, they recommend to supplement $SRMR$ by the *incremental* fit index CFI, defined as follows:

$$CFI = \begin{cases} 0, & \text{if } FI < 0 \\ FI & \\ 1, & \text{if } FI > 1, \end{cases} \tag{2.5.8}$$

where

$$FI = \frac{[(G_{\text{null}} - df_{\mathsf{null}}) - (G_{\text{h}} - df_{\mathsf{h}})]}{(G_{\text{null}} - df_{\mathsf{null}})},$$

where G_{null} and df_{null} are the test statistic G and the degrees of freedom, respectively, for the null model hypothesizing that the covariance matrix should be a diagonal matrix with free diagonal elements, while G_{h} and df_{h} are the G test statistic and the degrees of freedom, respectively, for the hypothesized model. As indicated by its property, the CFI represents the increase (increment) in model fit relative to a baseline model, which in the present case is the null modell. It was found by Hu and Bentler (1999) that a cutoff value for CFI should be close to 0.95. All these goodness-of-fit indices are easily obtained from the output, provided by the R package `sem`.

It is worth pointing out that it is recommended to use the goodness-of-fit indices for assessing the fit of a number of competing models fitted to the same data set, rather than the fit of a single model. Researches also should pay attention to other aspects of model fit such as examining parameter estimates to ensure they have the anticipated signs and magnitudes. Before

considering some type of model modification, other reasons why a model may not fit, such as small sample size, nonnormality, or missing data, need to be ruled out first (Boomsma, 2000).

An important comment concerning the `o.i.FA(2,1)`-model is that despite the overidentifiability, an exact ML solution for α_{f} in terms of the sample variances and covariances of the indicators does exist. This is because the model in its unstandardized form can be viewed as a ME model under the condition that both error variances are known:

$$\left\{\begin{array}{lllll} x_{\mathsf{f}\,t} = & \alpha_{\mathsf{f}} \cdot \xi_{\mathsf{f}\,t} & + & \delta_{\mathsf{f}\,t} & \\ z_t = & \xi_{\mathsf{f}\,t} & + & \underbrace{\gamma_t}_{=\eta_{\mathsf{internal}}+\epsilon_t} & . \end{array}\right. \tag{2.5.9}$$

As the model is overidentified, the method of moments approach cannot be used to derive the ML estimates. The estimators were obtained by maximizing the likelihood directly by Birch (1964). They are given by:

$$\widehat{\alpha}_{\mathsf{f}} = \frac{s^2_{x_{\mathsf{f}}} - \ell \cdot s^2_z + \sqrt{\left(s^2_{x_{\mathsf{f}}} - \ell \cdot s^2_z\right)^2 + 4 \cdot \ell \cdot (s_{x_{\mathsf{f}}z})^2}}{2 \cdot s_{x_{\mathsf{f}}z}}, \tag{2.5.10}$$

where $\ell = \sigma^2_{\delta_{\mathsf{f}}}/\sigma^2_{\gamma}$,

$$\widehat{\sigma}^2_{\xi_{\mathsf{f}}} = \frac{s^2_{x_{\mathsf{f}}} + \ell \cdot s^2_z - 2 \cdot \sigma^2_{\delta_{\mathsf{f}}} + \sqrt{(s^2_{x_{\mathsf{f}}} - \ell s^2_z)^2 + 4 \cdot \ell \cdot (s_{x_{\mathsf{f}}z})^2}}{2 \cdot (\ell + \widehat{\alpha}^2_{\mathsf{f}})}, \tag{2.5.11}$$

provided one or more of the following conditions is satisfied: $s^2_z > \sigma^2_{\gamma}$, $s^2_{x_{\mathsf{f}}} > \sigma^2_{\delta_{\mathsf{f}}}$ or $s^2_z > (\sigma^2_{\gamma} - s^2_z) \cdot (\sigma^2_{\delta_{\mathsf{f}}} - s^2_{x_{\mathsf{f}}})$. Estimator (2.5.10) is valid both for the functional and structural models. According to Cheng (Sec. 2.1.3), the asymptotic distribution of $\sqrt{n}(\widehat{\alpha}_{\mathsf{f}} - \alpha_{\mathsf{f}})$ tends to a normal distribution with zero mean and the variance

$$\Gamma = \frac{\sigma^2_{\gamma}}{\sigma^2_{\xi_{\mathsf{f}}}} \cdot (\ell + \alpha^2_{\mathsf{f}}) + \ell \cdot \left(\frac{\sigma^2_{\gamma}}{\sigma^2_{\xi_{\mathsf{f}}}}\right)^2, \tag{2.5.12}$$

obtained by means of the delta method. By replacing unknown parameters in (2.5.12) by their consistent estimates, an estimator of the variance of the limiting distribution of $\widehat{\alpha}_{\mathsf{f}}$ is obtained:

$$\widehat{\mathrm{Var}}(\widehat{\alpha}_{\mathsf{f}}) = n^{-1} \cdot \left(\frac{\sigma^2_{\gamma}}{\widehat{\sigma}^2_{\xi_{\mathsf{f}}}} \cdot (\ell + \widehat{\alpha}^2_{\mathsf{f}}) + \ell \cdot \left(\frac{\sigma^2_{\gamma}}{\widehat{\sigma}^2_{\xi_{\mathsf{f}}}}\right)^2\right). \tag{2.5.13}$$

Using the corresponding standard error, we may construct an approximate $1-p$ confidence interval for α_{f} according to (2.3.12), i.e. the Wald confidence interval. Hence, two different parameterizations of the `o.i.FA(2,1)`-model permit us to construct two confidence sets for $\lambda_{11}/\lambda_{21}$ (equal to α_{f} under both parameterizations):

1. The Wald confdence interval in (2.3.12) under the unstandardized parameterization of the model, i.e. under the `o.i.ME` model in (2.5.9);

2. The Fieller confidence set by solving inequality (2.3.25) under the standardized form of the model, i.e. under the `o.i.FA(2,1)`-model in (2.5.4).

In this work, we will compare the performance of both methods. From (2.5.10) we again may conclude that a sufficiently large covariance between the simulated temperature x_{f} and the proxy z is an important premize for obtaining a reasonable estimate of the amplitude of a forcing effect. Besides, as follows from (2.5.13), a sufficiently large $\sigma^2_{\xi_{\mathsf{f}}}$ is also needed for ensuring reasonable confidence intervals. Once again, we are faced with the issue of testing $H_0 : \sigma^2_{\xi_{\mathsf{f}}} = 0$, which is not so easy to do under the ME model specification, but under the factor model specification the test of the equivalent hypothesis $H_0 : \lambda_{21} = \sigma^2_{\xi_{\mathsf{f}}} = 0$ is available.

2.5.2 Approach 2: $\mathbf{Var}(\xi_{\mathsf{f}}) = 1$

By replacing τ in the FA(2,2)-model in (2.4.13) by the proxy z with a time-constant precision and applying the second identification approach to it, we obtain:

$$\begin{cases} x_{\mathsf{f}\,t} = & \alpha_{\mathsf{f}} \cdot \xi_{\mathsf{f}\,t} + & 0 \cdot \xi_{\mathsf{total}\perp\mathsf{f}} & + & \delta_{\mathsf{f}\,t} \\ z_t = & \kappa \cdot \xi_{\mathsf{f}\,t} + & \xi_{\mathsf{total}\perp\mathsf{f}} & + & \underbrace{\gamma_t}_{\eta_{\mathsf{internal}\;t}+\epsilon_t} \end{cases}, \tag{2.5.14}$$

where the distributional assumptions of the involved variables are the same as under model (2.5.1), obtained under Approach 1, with an additional assumption that $\sigma^2_{\xi_{\mathsf{f}}} = 1$. Regarding the variance of the second latent factor, two cases are considered: (1) $\mathrm{Var}(\xi_{\mathsf{total}\perp\mathsf{f}}) \neq 1$, and (2) $\mathrm{Var}(\xi_{\mathsf{total}\perp\mathsf{f}}) = 1$

`Case 1:` $\mathrm{Var}(\xi_{\mathsf{total}\perp\mathsf{f}}) \neq 1$
Reparameterizing the model above in the similar manner as model (2.5.1),

i.e.

$$\begin{cases} x_{\mathsf{f}\,t} = & \lambda_{11} \cdot \xi_{\mathsf{f}\,t} + & 0 & \cdot \xi'_{\mathsf{total}\perp\mathsf{f}} & + & \delta_{\mathsf{f}\,t} \\ z_t = & \lambda_{21} \cdot \xi_{\mathsf{f}\,t} + \lambda_{22} \cdot & \xi'_{\mathsf{total}\perp\mathsf{f}} & & + & \underbrace{\gamma_t}_{\eta_{\mathsf{internal}\ t} + \epsilon_t} \end{cases}, \tag{2.5.15}$$

$$\text{where} \quad \lambda_{11} = \alpha_{\mathsf{f}}, \ \ \lambda_{21} = \kappa, \ \text{and} \ \lambda_{22} = \sigma_{\xi_{\mathsf{total}\perp\mathsf{f}}},$$

shows clearly the equivalence between this model and the standardized `j.i.FA(2,2)`-model in (2.5.2). In particular, the estimate of $\lambda_{11}/\lambda_{21}$ under both models is the same. Moreover, the confidence set for $\lambda_{11}/\lambda_{21}$ is to be constructed according to the same method, namely the Fieller method.

`Case 2:` $\mathrm{Var}(\xi_{\mathsf{total}\perp\mathsf{f}}) = 1$
When fixing the variance of $\xi_{\mathsf{total}\perp\mathsf{f}}$ to one, we obtain a model that says that when this latent factor changes 1 unit, the second indicator variable changes by 1 unit[13], given that the first latent factor is held fixed. That is a very strong and unrealistic hypothesis, requiring the variance of z to be at least 1. This makes us to refrain from taking this model into consideration.

To conclude, since the standardized `j.i.FA(2,2)`-model under Approach 2 does not provide a new estimator of $\lambda_{11}/\lambda_{21}$ and new methods of calculating the associated confidence set, it is sufficient to consider the standardized `j.i.FA(2,2)`-model only under Approach 1.

Before moving on to the discussion about the model estimation in the presence of heteroscedasticity, let us take a closer look at a situation when the estimate of σ_γ^2 is not appropriate, i.e. $\widehat{\sigma}_\gamma^2 > s_z^2$. In that case, the model cannot be estimated. Admittedly, even the estimate of the internal variability of a climate model under consideration, $\sigma_{\delta_{\mathsf{f}}}^2$, might turn out to be larger than $s_{x_{\mathsf{f}}}^2$, making the model underidentifiable, but it is definitely more challenging to obtain a precise estimate of σ_γ^2 than a precise estimate of $\sigma_{\delta_{\mathsf{f}}}^2$. Should it happen that $\widehat{\sigma}_\gamma^2 > s_z^2$, the parameter σ_γ^2 ought to be treated as a free parameter. Obviously, this requirement makes the `j.i.FA(2,2)`-model underidentifiable. In order to permit identifiability more than two indicators should be incorporated. This was our main motive of proposing a three-indicator model, described in the next section.

[13]Note that when temperature data is analyzed, it is assumed that latent factors have the same unit of measurement as observed variables. By scientific convention the scale for temperature is measured in Kelvin units.

2.5.3 Extension of two-factor models. Approach 1: $\kappa = 1$

Suppose that a climate model simulation influenced by all external (but of course reconstructed) forcings, including a forcing f that has influenced x_{f}, is available. Let this simulation be denoted $\{x_{\mathsf{total}}\}$. The idea is to invoke the variable x_{total} as an indicator for the (true) latent variables ξ_{f} and $\xi_{\mathsf{total}\perp\mathsf{f}}$. By adding the *mean-centered* variable x_{total} to the standardized `j.i.FA(2,2)`-model, the following standardized three-indicator two-factor model, FA(3,2)-model, is obtained:

$$\begin{cases} x_{\mathsf{f}\,t} &= \underbrace{\alpha_{\mathsf{f}} \cdot \sigma_{\xi_{\mathsf{f}}}}_{=\lambda_{11}} \cdot \xi'_{\mathsf{f}\,t} + \quad 0 \cdot \ \xi'_{\mathsf{total}\perp\mathsf{f}} + \quad \delta_{\mathsf{f}\,t} \\ z_t &= \underbrace{\sigma_{\xi_{\mathsf{f}}}}_{=\lambda_{21}} \cdot \xi'_{\mathsf{f}\,t} + \underbrace{\sigma_{\xi_{\mathsf{total}\perp\mathsf{f}}}}_{=\lambda_{22}} \cdot \ \xi'_{\mathsf{total}\perp\mathsf{f}} + \underbrace{\gamma_t}_{\eta_{\mathsf{internal}\ t}+\epsilon_t} \\ x_{\mathsf{total}\ t} &= \underbrace{\kappa_1 \cdot \sigma_{\xi_{\mathsf{f}}}}_{=\lambda_{31}} \cdot \xi'_{\mathsf{f}\,t} + \underbrace{\kappa_2 \cdot \sigma_{\xi_{\mathsf{total}\perp\mathsf{f}}}}_{=\lambda_{32}} \cdot \xi'_{\mathsf{total}\perp\mathsf{f}} + \quad \delta_{\mathsf{total}\ t}, \end{cases} \tag{2.5.16}$$

where the variables $\delta_{\mathsf{f}\,t}$, γ_t and $\delta_{\mathsf{total}\ t}$ are mutually independent normally distributed variables with zero mean and variances $\sigma^2_{\delta_{\mathsf{f}}}$, σ^2_γ and $\sigma^2_{\delta_{\mathsf{total}}}$, respectively. The parameter of interest is the ratio $\lambda_{11}/\lambda_{21}$, which under model (2.5.16) is equal to α_{f}, as it should be under Approach 1. Besides that, model (2.5.16) provides an opportunity to estimate the amplitude of a forcing effect of a forcing f in x_{total}. That is, another parameter of interest is the ratio $\lambda_{31}/\lambda_{21}$.

The model's reproduced variance-covariance matrix is:

$$\begin{aligned} \sigma^2_{x_{\mathsf{f}}} &= \lambda^2_{11} + \sigma^2_{\delta_{\mathsf{f}}} \\ \sigma_{x_{\mathsf{f}} z} &= \lambda_{11} \cdot \lambda_{21} \\ \sigma_{x_{\mathsf{f}} x_{\mathsf{total}}} &= \lambda_{11} \cdot \lambda_{31} \\ \sigma^2_z &= \lambda^2_{21} + \lambda^2_{22} + \sigma^2_\gamma \\ \sigma^2_{z x_{\mathsf{total}}} &= \lambda_{21} \cdot \lambda_{31} + \lambda_{22} \cdot \lambda_{32} \\ \sigma^2_{x_{\mathsf{total}}} &= \lambda^2_{31} + \lambda^2_{32} + \sigma^2_{\delta_{\mathsf{total}}}\,. \end{aligned} \tag{2.5.17}$$

Since (2.5.17) contains six unique elements, at most six parameters can be estimated. That is, a vector of free parameters for an just-identified model includes all five factor loadings and one of the specific variances. As there are three specific variances, three different just-identified models can be formulated. Which specific variance that should be treated as free depends on

such factors as availability of replicates of a corresponding climate model or on how confident a researcher is in known values of the specific variances. As discussed earlier, it might be of more interest to treat the specific variance of γ as free. In that case, $\sigma^2_{\delta_{\mathsf{f}}}$ and $\sigma^2_{\delta_{\mathsf{total}}}$ are treated as known and estimated directly from replicates of the corresponding climate models.

As known, just-identified factor models are associated with explicit expressions of model estimates in terms of the sample variances and covariances of indicators. However, we do not provide neither estimators nor associated side conditions for a proper solution for any just-identified model of those three possible. It is supposed that the model of interest is estimated by means of a specialized software package, e.g. the *R* package `sem`, used in this work. If all side conditions for a proper solution are fulfilled, a program provides an output with the parameter estimates with the associated matrix of the estimated variances and covariances among the estimates. Based on this information, researches can calculate an estimate of $\lambda_{11}/\lambda_{21}$ with the associated confidence set, constructed according to the Fieller method, i.e. by solving inequality (2.3.25). If no solution is obtained, which occurs if the information matrix is singular, or if an improper solution, e.g. Heywood case, is obtained, the model should be reformulated accordingly.

We do not exclude a situation when all specific variances are regarded as known, which implies that only factor loadings are to be estimated. The resulting model is overidentified with one degree of freedom, which allows us to assess its overall fit to the data. Provided the solution is proper and interpretable (in terms of both sign and size), the overall fit is assessed statistically by the G statistic, defined in (2.3.20) and heuristically using a number of goodness-of-fit indices, e.g. the indices defined in $(2.5.5. - 2.5.8)$. The result of the test and the values of the indices can be easily retrieved from the output (see Appendix for an example of fitting the overidentified FA(3,2)-model).

Further, both just-identified models and the overidentified model above can be simplified by fixing the loadings λ_{22} and/or λ_{32} to zero. An importnat point to realize about the process of model simplification is that even if simplification is justified from the climatological point of view and supported empirically by insignificant values of the relevant estimates, fixing one or two loadings to a specified value might lead to unpredictable effects on the estimates of the remaining parameters both in terms of sign and magnitude that perhaps cannot be linked to climatological properties of the involved variables. That is, the process of model simplification should be performed with caution.

2.5.4 Extension of two-factor models. Approach 2: $\mathbf{Var}(\xi_{\mathsf{f}}) = 1$

Just as with the standardized `j.i.FA(2,2)`-model associated with Approach 1 (see Sec. 2.5.2)), it is sufficient to consider only the standardized FA(3,2)-model (2.5.16) under Approach 1.

2.6 Models with two latent factors. Heteroscedasticity

When fitting the two-factor models (the `j.i.FA(2,2)`-model or the FA(3,2)-model which can be either just-identified or overidentified) to data containing climate observations with time-varying precision, one can proceed precisely in the same way as for the one-factor models, i.e. by analyzing the weighted sample variance-covariance matrix of the indicator instead of the ordinary one. Under the two-factor models, these matrices are calculated as follows:

under the `j.i.FA(2,2)`-model:

$$\begin{bmatrix} s^2_{x_{\mathsf{f}}} & s^{(w)}_{x_{\mathsf{f}}z} \\ s^{(w)}_{x_{\mathsf{f}}z} & s^{2(w)}_{z} \end{bmatrix}$$

under the `FA(3,p)`-model, $p = 1, 2$:

$$\begin{bmatrix} s^2_{x_{\mathsf{f}}} & s^{(w)}_{x_{\mathsf{f}}z} & s_{x_{\mathsf{f}}x_{\mathsf{total}}} \\ s^{(w)}_{x_{\mathsf{f}}z} & s^{2(w)}_{z} & s^{(w)}_{zx_{\mathsf{total}}} \\ s_{x_{\mathsf{f}}x_{\mathsf{total}}} & s^{(w)}_{zx_{\mathsf{total}}} & s^2_{x_{\mathsf{total}}}, \end{bmatrix} \tag{2.6.1}$$

where the unique elements are:

$$s^2_{x_{\mathsf{f}}} = \frac{\sum_t (x_{\mathsf{f}\,t} - \bar{x}_{\mathsf{f}})^2}{n}, \qquad \text{where} \quad \bar{x}_{\mathsf{f}} = \frac{\sum_t x_{\mathsf{f}}}{n},$$

$$s^{(w)}_{x_{\mathsf{f}}z} = \frac{\sum_t w_t \left(x_{\mathsf{f}\,t} - \bar{x}^{(w)}_{\mathsf{f}}\right)\left(z_t - \bar{z}^{(w)}\right)}{\sum_t w_t}, \qquad \text{where} \quad \bar{x}^{(w)}_{\mathsf{f}} = \frac{\sum_t w_t x_{\mathsf{f}}}{\sum_t w_t}, \quad \text{etc.}$$

$$s_{x_{\mathsf{f}}x_{\mathsf{total}}} = \frac{\sum_t (x_{\mathsf{f}\,t} - \bar{x}_{\mathsf{f}})(x_{\mathsf{f}\,t} - \bar{x}_{\mathsf{total}})}{n},$$

$$s^{2(w)}_{z} = \frac{\sum_t w_t^2 \left(z_t - \bar{z}^{(w)}\right)^2}{\sum_t w_t^2}, \qquad \text{where} \quad \bar{z}^{(w)} = \frac{\sum_t w_t^2 z_t}{\sum_t w_t^2},$$

$$s^{(w)}_{zx_{\mathsf{total}}} = \frac{\sum_t w_t \left(z_t - \bar{z}^{(w)}\right)\left(x_{\mathsf{total}\,t} - \bar{x}^{(w)}_{\mathsf{total}}\right)}{\sum_t w_t},$$

$$s^2_{x_{\mathsf{total}}} = \frac{\sum_t (x_{\mathsf{total}\ t} - \bar{x}_{\mathsf{total}})^2}{n},$$

where the weights w_t are defined in (2.4.12), i.e.

$$w_t = \frac{\sigma^2_{\eta_{\mathsf{internal}}}}{\sigma^2_{\eta_{\mathsf{internal}}} + \sigma^2_\epsilon(t)} = \frac{\sigma^2_{\eta_{\mathsf{internal}}}}{\sigma^2_\gamma(t)}.$$

Reasoning as in (2.4.5), it was found that the parameters estimated by the weighted *covariances* are the same as the parameters estimated by the ordinary ones, for example

$$E\left[s^{(w)}_{x_{\mathsf{f}}\, z}\right] = E\left[\frac{\sum_{t=1}^n w_t \left(x_{\mathsf{f}\ t} - \mu_{x_{\mathsf{f}}}\right)(z_t - \mu_z)}{\sum_{t=1}^n w_t}\right] = \sigma_{x_{\mathsf{f}} z} = E\left[s_{x_{\mathsf{f}}\, z}\right] = \lambda_{11} \cdot \lambda_{21}.$$

On the other hand, the expected value of the weighted variance of z,

$$E\left[s^{2(w)}_z\right] = E\left[\frac{\sum_{t=1}^n w_t^2 \left(z_t - \mu_z\right)^2}{\sum_{t=1}^n w_t^2}\right] = \lambda^2_{21} + \lambda^2_{22} + \sigma^{2(w)}_\gamma,$$

includes a new parameter: the unknown weighted average variance of the noise term γ over the entire time period analyzed, $\sigma^{2(w)}_\gamma = \sum_t w_t^2 \sigma^2_\gamma(t) \Big/ \sum_t w_t^2$. However, with the a priori known weights in hand, it is not difficult to calculate the value of $\sigma^{2(w)}_\gamma$, thereby allowing to treat this parameter as known. A conceivable problem that might occur in practice is similar to that arising under homoscedasticity, i.e. it might happen that $\sigma^{2(w)}_\gamma > s^{2(w)}_z$. As a resort, $\sigma^{2(w)}_\gamma$ could be treated as a free parameter. Unfortunately, the `j.i.FA(2,2)`-model is not estimable in that case, but the `j.i.FA(3,2)`-model with all factor loadings and $\sigma^{2(w)}_\gamma$ as model parameters is. Here we have nevertheless to bear in mind that the weights, used for calculating the weighted sample variance-covariance matrix, are not precise.

2.7 Usage of mean time series

It is not uncommon in climatological statistic to analyze ensemble-mean sequences instead of single members of an ensemble. Averaging over replicates of the same type of forced model leads to a time series with an enhanced forced climate signal and a reduced effect of the internal variability of the corresponding forced climate model. It is, of course, possible to use mean time series belonging to an ensemble even in our own analysis.

Assuming that k and g replicates of the x_{f}- respective x_{total}-climate models are available, the single simulations $\{x_{\mathsf{f}\,t}\}$ and $\{x_{\mathsf{total}\,t}\}$ can be replaced by the mean time series $\{\bar{x}_{\mathsf{f}\,t}\}$ and $\{\bar{x}_{\mathsf{total}\,t}\}$, respectively, obtained by averaging over k respective g replicates at each time point t. This entails that the specific variances in our statistical models, $\sigma^2_{\delta_{\mathsf{f}}}$ and $\sigma^2_{\delta_{\mathsf{total}}}$, should be replaced by $\sigma^2_{\delta_{\mathsf{f}}}/k$ and $\sigma^2_{\delta_{\mathsf{total}}}/g$, respectively, where $\sigma^2_{\delta_{\mathsf{f}}}$ and $\sigma^2_{\delta_{\mathsf{total}}}$ are calculated by applying estimator (2.3.3). Naturally, the sample variances and covariances involving the single sequences $\{x_{\mathsf{f}\,t}\}$ and $\{x_{\mathsf{total}\,t}\}$ should be replaced by the sample variance and covariances involving their averaged counterparts.

2.8 Summary

Before proceeding further with a numerical analysis, let us summarize briefly our theoretical discussion.

Several factor models have been suggested as appropriate models for estimating the amplitude of a forcing effect in a climate model. Examination of the structure of the expression for the true temperatue τ in the basic statistical model (2.1) revealed that not only one-factor models, but also two-factor models can be useful. Initially, a `just-identified` two-indicator one-factor model was formulated:

`j.i.FA(2,1)`-model:

$$\begin{cases} x_{\mathsf{f}\,t} = \alpha_{\mathsf{f}} \cdot \xi_{\mathsf{f}\,t} \; + & \delta_{\mathsf{f}\,t} \\ z_t \;\; = \kappa \cdot \xi_{\mathsf{f}\,t} \; + & \underbrace{\nu_t}_{=\xi_{\mathsf{total}\perp\mathsf{f}\,t}+\eta_{\mathsf{internal}\,t}+\epsilon_t} \end{cases}$$

By applying two different identification approaches, we could either reformulate this model as a `j.i.ME` model (see (2.3.1)), or keep its original form as a `j.i.FA(2,1)`-model (see 2.3.14)). More presicely, the `j.i.ME`-model was obtained under Approach 1, assuming $\kappa = 1$, while the `j.i.FA(2,1)`-model under Approach 2, assuming $\mathrm{Var}(\xi_{\mathsf{f}}) = 1$.

Regardless of approach, the estimator of the amplitude of a forcing effect under both models is the same as well as the side conditions for a proper/admissible solution. The sole difference is that the amplitude of a forcing effect is represented by the parameter α_{f} under the ME model and by the ratio of the factor loadings $\alpha_{\mathsf{f}}/\kappa$ under the `j.i.FA(2,1)`-model. Thus, two different methods of constructing a confidence set for the amplitude of a forcing effect can be employed. Under the former model, an approximate $1-p$ large sample (always bounded) confidence interval, known as the Wald confidence interval, is constructed (see 2.3.12). Under the latter, a confidence set is

constructed on the basis of the Fieller method of finding the confidence set of the ratio of two normal means, which in the context of our analysis corresponds to solving the quadratic inequality in (2.3.25). The Fieller confidence set does not suffer from confidence level errors as does the Wald confidence interval.

Moving the orthogonal complement $\xi_{\mathsf{total}\perp\mathsf{f}\,t}$ from the specific factor ν_t and regarding it as a second latent factor, uncorrelated with the first latent factor, $\xi_{\mathsf{f}\,t}$, made it possible to formulate (orthogonal) *two-factor* models. These models differ in the number of indicators, but all of them, when Approach 1 is applied, are subject to reparameterization based on the standardization of the latent factors. Thanks to reparameterization, the amplitude of a forcing effect can be represented by the ratio of two factor loadings, namely $\lambda_{11}/\lambda_{21}$, where $\lambda_{11} = \alpha_{\mathsf{f}} \cdot \sigma_{\xi_{\mathsf{f}}}$ and $\lambda_{21} = \sigma_{\xi_{\mathsf{f}}}$, making it possible to construct the Fieller confidence set for α_{f} even under Approach 1. In fact, a standardized factor model under Approach 1 is equivalent to a factor model obtained under Approach 2. Despite this equivalence, the focus is on the former type because rewriting some factor models in their unstandardized form can reveal interesting links to ME models.

Regarding two-indicator models, the initial model is the `j.i.FA(2,2)`-model in (2.5.2). The model is identical to the `j.i.FA(2,1)`-model in terms of the estimator of the amplitude and the associated confidence set. However, in contrast to the `j.i.FA(2,1)`-model, the `j.i.FA(2,2)`-model can be modified by eliminating the second latent factor, which results in an overidentified two-indicator one-factor model, associated with a new estimator of the amplitude of a forcing effect and a new Fieller confidence set for it. The model is defined in (2.5.4) and its abbreviation is `o.i.FA(2,1)`.

A distinguishing feature of this model is that in its unstandardized form the model is actually an (overidentified) ME model, abbr. `o.i.ME`. The estimator of the amplitude remains of course the same as under the `o.i.FA(2,1)`-model, but the associated confidence set is the Wald confidence interval. Once again, we are observing the situation when one estimator of the amplitude is associated with different methods of constructing a confidence set for it.

The initial three-indicator model is the `FA(3,2)`-model in (2.5.16). Its main advantage is that it offers flexibility in choosing known parameters. Depending on what parameters are chosen and on their number, various just-identified as well as various overidentified three-indicator models can be formulated. In Table 1 in Sec. 3.3, one can see models that are suitable for our analysis.

The next stage of our analysis will be devoted to the numerical compar-

ison of the performance of the suggested models.

3 Numerical experiment to compare the statistical models

3.1 Description of the pseudo-proxy experiment

A first step in designing a numerical experiment is to determine criteria in accordance with the aim of a particular study. Our aim is to compare the performance of different estimators of the amplitude of a forcing effect in a climate model simulation. It is clear that without knowing the correct value of the parameter, representing the amplitude of a forcing effect, that is $\lambda_{11}/\lambda_{21}$, it is not possible to make a comparison between different estimators. As stated in the basic statistical model (2.1), a correct representation of the forcing effect in the climate model x_{f} corresponds to $\alpha_{\mathsf{f}} = 1$. This leads to the question: Against what data should x_{f} be analyzed in order to argue that the correct value of $\lambda_{11}/\lambda_{21}$ is 1?

Clearly, when analyzing x_{f} against real-world climate observations, uncertainties in the reconstruction of forcing f do not allow us to determine to what extend the response to a reconstructed forcing f differ from the response to its real-world counterpart. That is, the correct value of the parameter is not known.

Data for which the correct value of $\lambda_{11}/\lambda_{21}$ is not only known but also is equal to 1 are climate model simulations whose forcing history contains the *same* underlying forcing f, used for generation of x_{f}. In other words, real-world proxy data should be represented by climate model simulations. Suitable candidates for pseudo-proxy data are (1) a replicate of x_{f} or (2) any other simulation driven by a group of forcings containing the forcing f. In the former case, the orthogonal complement, $\xi_{\mathsf{total}\perp\mathsf{f}}$, does not exist, while in the latter case it does. Since the two-factor models involve the orthogonal complement as a second latent factor, it is logical for our analysis to represent climate observations by climate model simulations with a larger forcing history, i.e. a history built by adding various forcings to f.

Actually, we let these specially selected climate model simulations play the role of τ, for which, as known, $\sigma^2_\epsilon(t)$ is zero for all time points. The key point in it is that we can form a pseudo-proxy series $\{z_t\}$ by adding sequences of simulated values of $\epsilon_t \sim \mathrm{N}\left(0, \sigma^2_\epsilon(t) > 0\right)$ to the pseudo-τ. This in turn enable us (1) to control uncertainty sources associated with climate observations, which is not possible when real-world data are analyzed, and (2)

to investigate how sensitive the estimators of $\lambda_{11}/\lambda_{21}$ are to various levels of the noise in the pseudo-proxy. The method of constructing pseudo-proxies will be described in detail later in Sec. 4.

We would like to point out that the idea to use climate model simulations instead of real-world data is not new in climate statistics. Such experiments are known as *pseudo-proxy experiments*, abbr. `PPE`. An example on the use of `PPE` is the experiment that aims to test climate reconstruction methods (for its description see Smerdon, 2011). In this experiment, let us refer to it as `PPE-RecM`, the true temperature τ is also represented by climate model simulations. However, in contrast to our experiment, where the choice of pseudo-τ depends on x_{f}, `PPE-RecM` does not impose such restrictions on simulations. Further, since `PPE-RecM` aims to test reconstruction methods, the pseudo-τ is explicitly involved in calculating various metrics measuring the discrepancy between the pseudo-τ itself and the reconstructed temperature (during the reconstruction period). The aim of our own analysis is to compare performance of different estimators of the amplitude of a forcing effect in a climate model. It gives rise to other criteria that do not involve the pseudo-τ itself, for example the deviation of the estimates of $\lambda_{11}/\lambda_{21}$ from the correct value of the parameter, how reasonable the associated confidence sets are, or how well the overidentified factor models fit the data.

Further, just as in our own pseudo-proxy experiment, `PPE-RecM` also presumes the construction of pseudo-proxies, permitting researchers to study the sensitivity of statistical methods to increasing noise level. Nevertheless, in our `PPE`, misleading conclusions are possible. To avoid them, an appropriate hierarchical analysis of iteratively obtained estimates is needed for any level of proxy noise (this issue will be discussed later in Sec. 4).

A possible weakness of our `PPE` is its dependence on replicates of climate models involved. Replicates are needed first of all for the estimation of the internal variability of a corresponding climate model. The more replicates, the better. It will certainly yield a more precise estimate of the internal variability of climate models, which is highly desirable for obtaining proper solutions of the models parameters, in particular of $\lambda_{11}/\lambda_{21}$. In case replicates are not available, it is possible to estimate the internal variability of forced climate models by means of replicates of an unforced climate model, which are more offen available. Yet, this cannot guarantee that side conditions for a proper estimate of $\lambda_{11}/\lambda_{21}$ will be fulfilled. Should that happen, other methods of generating synthetic data could be employed, for instance, having generated n (temporally dependent or independent) values of the unobservable variables, possessing in that case known distributions, a set of observed data can be formed in accordance with a statistical model under

consideration.

In sum, having evaluated different aspects of our PPE, we may conclude that its application is motivated and feasible within our analysis, making us capable to address questions posed in the present work.

3.2 Description of data, its initial analysis and preliminaries

Data used in the present analysis were generated during the COSMOS Millennium Activity simulation experiments conducted using the Max Planck Institute Earth System Model (MPI-ESM), representing climate conditions essentially within the last millennium. A detailed description of the model and the Millennium experiment can be found in Jungclaus et al. (2010). Below, the analyzed x-sequences with their notations and brief descriptions are listed:

1. $\{x_{\mathsf{unforced},t}\}$, a single unforced control simulation, spanning a period of 3000 years. This simulation was run under 800 AD orbital conditions and constant preindustrial greenhouse gas concentrations. For the purposes of this analysis, the control simulation has been separated into three 1000-yr long series that provides us with three replicates of x_{unforced}.

2. $\{x_{\mathsf{land\ use},t}\}$ is a result of adding to the control boundary conditions a reconstruction of temporal and spatial changes in *a land use* forcing, a forcing based on the civilized world's agricultural effect on the land surface and how that affects climate, e.g. through changed solar reflectivity. The sequence starts at 800 and runs until 2005, and it is a single simulation;

3. $\{x_{\mathsf{volcanic},t}\}$ was obtained by adding to the control boundary conditions a reconstruction of another single forcing, namely *volcanic* forcing caused by the estimated effect of past volcanic eruptions. The sequence spans the time 800 to 2005 AD, and it is a single simulation;

4. $\{x_{\mathsf{E1},t}\}$ is a multi-forcing simulation obtained by driving the climate model with changing *orbital* forcing conditions from 800 to 2005 AD, reconstructions of natural and anthropogenic *greenhouse gas and aerosol* forcing, together with the same as for $x_{\mathsf{land\ use}}$ reconstruction of *land use* forcing, the same as for x_{volcanic} reconstruction of *volcanic* forcing in addition to a reconstruction of *solar* forcing with a low amplitude

(with an increase in total solar irradiance of 0.1% from the Maunder Minimum at 1647-1715 AD until the present). In total, there are five sequences, i.e. 5 replicates of x_{E1}, forming the 'E1' ensemble.

5. $\{x_{\mathsf{E2},t}\}$, is another multi-forcing simulation influenced by same forcings as for E1, but with another solar forcing having a high amplitude (with an increase from the Maunder Minimum by 0.25%). In total, there are three sequences, i.e. 3 replicates of x_{E2}, forming the 'E2' ensemble.

All above data are originally available as monthly averages at the climate model grid resolution of 3.75°, from which the desired spatial and temporal averages can be calculated. The larger the spatial scale, the stronger the forcing-related component of temperature variation becomes as compared to the internal variability. Therefore, the global-mean land-only temperature is chosen for our analysis[14]. The time series have been transformed from monthly series to annual mean time series. By doing so we remove the seasonal periodicity typical for monthly data.

The time period analyzed in the present work is the 1000-year long period $850 - 1849$ AD. The industrial period after 1850 AD has been omitted in order to eliminate complicating influence of the anthropogenic greenhouse gas emissions.

The role of the x_{f}-sequence will be played by each of two single forcing simulations, namely x_{volcanic} and $x_{\mathsf{land\text{-}use}}$. This choice is motivated by the opposite properties that these two forcings are expected to have at the temporal- and spatial scales analyzed. More specifically, the `volcanic` forcing is expected to cause substantial temperature changes, while the `land use` forcing is not. In terms of our statistical models, it means that the hypothesis $H_0 : \sigma^2_{\xi_{\mathsf{f}}} = 0$ is expected to be rejected for the `volcanic` forcing, but not for the `land use` forcing. Consequently, significant respective insignificant values of the U_R-statistic for these two single forcing simulations are expected as well.

The true temperature τ will be represented by each replicate of x_{E1} and x_{E2}. Each of these replicates will also be invoked as x_{total}, needed as a third indicator in the FA(3,2)-model in (2.5.16), but such that $\tau \neq x_{\mathsf{total}}$. Recall from Sec. 3.1 that replacing the true temperature τ by climate model simulations with a larger forcing history such that the reconstruction of the forcing f that influenced x_{f} in included, leads to: (1) the orthogonal complement $\xi_{\mathsf{total}\perp\mathsf{f}}$ exists and (2) the correct value of $\lambda_{11}/\lambda_{21}$ is 1. Note the the latter statement is true regardless of whether the hypothesis $H_0 : \sigma^2_{\xi_{\mathsf{f}}} = 0$ is rejected or not.

Another consequence of replacing τ by a climate model simulation is

[14]Data used here are the same as in Hind et al., 2012. Their motivation to exclude ocean temperatures was that most real temperature proxy data come from land regions.

that η_{internal} corresponds now to the internal variation of a climate model. In particular, letting τ be represented by a member from the E1 or E2 ensembles means that the variable η_{internal} corresponds to the variables δ_{E1} and δ_{E2} , respectively. Consequently, the specific factors in our two-factor models under the assumption that $\epsilon_t = 0$ are to be represented by δ_{f}, δ_{E1} and/or δ_{E2}, which is illustrated and exemplified in (3.2.1):

$$\begin{cases} x_{\mathsf{f}\,t} & = & \lambda_{11}\cdot\,\xi'_{\mathsf{f}\,t} & + & 0\cdot\,\xi'_{\mathsf{total}\perp\mathsf{f}} & + & \delta_{\mathsf{f}\,t} \\ \underbrace{\tau_t}_{\equiv x_{\mathsf{E1,\ repl.i}\ t}} & = & \lambda_{21}\cdot\,\xi'_{\mathsf{f}\,t} & + & \lambda_{22}\cdot\,\xi'_{\mathsf{total}\perp\mathsf{f}} & + & \underbrace{\eta_{\mathsf{internal}}}_{\equiv\delta_{\mathsf{E1,\ repl.i}\ t}} \\ \underbrace{x_{\mathsf{total}\ t}}_{=x_{\mathsf{E2,\ repl.\ j}\ t}} & = & \lambda_{31}\cdot\,\xi'_{\mathsf{f}\,t} & + & \lambda_{32}\cdot\,\xi'_{\mathsf{total}\perp\mathsf{f}} & + & \underbrace{\delta_{\mathsf{total}\ t}}_{=\delta_{\mathsf{E2,\ repl.j}\ t}} \end{cases},$$

where $i = 1, 2, 3, 4, 5,\ j = 1, 2, 3.$

Since the ML estimates of the model parameters are obtained under the normality and independence assumptions, we need to investigate whether these assumptions are satisfied. First of all, it is of importance to investigate whether these assumptions are satisfied for each $\{\delta_t\}$-sequence. Regarding the latent-factor variables, it is possible to consider them as fixed unknown constants.

Let us first investigate whether the independence assumption is satisfied. Since the specific factors are unobservable, the series to analyze are:

$$\{x_{\mathsf{repl.i}\ t} - \bar{x}_{.t}\}, \quad i = 1, 2, \ldots, k,$$

depending obviously only on δ:s. As for δ_{E1} and δ_{E2}, the analysis of the autocorrelation structure of

$$\{x_{\mathsf{E1,\ repl.i}\ t} - \bar{x}_{\mathsf{E1}.t}\}, \quad i = 1, 2, 3, 4, 5$$

and

$$\{x_{\mathsf{E2,\ repl.i}\ t} - \bar{x}_{\mathsf{E2}.t}\}, \quad i = 1, 2, 3,$$

showed that all series exhibit a significant autocorrelation on the annual scale (see the left upper plot in Figure A1 in Appendix). Therefore, temporal aggregation of each time series by taking m-yr nonoverlapping averages has been performed for several values on m. The analysis of the autocorrelation functions of the new series revealed that the smallest time unit appropriate to apply to both ensembles is $m = 10$. The corresponding subplot in Figure A1 shows that the autocorrelation coefficients for all lags are insignificant as they are falling within the confidence bounds.

Regarding $\delta_{\mathsf{volcanic}}$ and $\delta_{\mathsf{land\ use}}$, we unfortunately cannot perform the similar analysis, because the *COSMOS* experiment did not involve any replicates of the single forcing simulations. What we can do is to assume that the decadal $\{\delta_{\mathsf{volcanic}\ t}\}$- and $\{\delta_{\mathsf{land\ use}\ t}\}$-sequences do not exhibit a significant autocorrelation, which does not seem to be an unreasonable assumption. Hence, taking 10-year non-overlapping means in all x-sequences reduces number of observations in each of them from 1000 to 100.

An important point to realize about time aggregation of data, performed in this way, is that it does not necessarily lead to an insignificant autocorrelation of x-sequences themselves. A natural explanation for this phenomenon is the presence of the forcing-related component - ξ_{f} in the single forcing simulations and ξ_{total} in the multiforcing simulations - which may exhibit a significant autocorrelation even after time-aggregation of data. For instance, the analysis of x_{E1}-and x_{E2}-sequences showed that most of them display a significant autocorrelation even for $m = 20$ (see Figure A2 in Appendix). For the present analysis, however, time units of 20 or more years were not applied to avoid the effect of small sample sizes on the estimation procedure.

The presence of the significant autocorrelation in the multi-forcing simulations themselves leads to the conjecture that under the ME model in (2.3.1) and the `j.i.FA(2,1)`-model in (2.3.14), a significant autocorrelation may be introduced into the specific factor η, or ν when $\sigma_\epsilon^2 > 0$, because the orthogonal complement $\xi_{\mathsf{total}\perp\mathsf{f}}$ is regarded as a part of it. It is clear that we do not have any possibility to check for the absence of autocorrelation in the orthogonal complement, therefore we have to assume a negligible autocorrelation in the $\{\xi_{\mathsf{total}\perp\mathsf{f}\ t}\}$-sequences in order to satisfy the independence assumption among the errors η_t. For the two-factor models, this conjecture is redundant because, as mentioned earlier, both $\xi_{\mathsf{f}\ t}$ and $\xi_{\mathsf{total}\perp\mathsf{f}\ t}$ can be regarded as fixed unknown constants for all t.

Having determined the appropriate time unit, we can now look for evidence of non-normality in the time-aggregated δ-depending series and the time-aggregated x_{volcanic} - and $x_{\mathsf{land\ use}}$-sequences themselves. It was chosen to estimate a density function for each series. It was also possible to apply a formal test for normality such as the Shapiro test or Kolmogorov-Smirnov test to the δ-depending sequences, but not to the x_{volcanic}- and $x_{\mathsf{land\ use}}$ - simulations. This is because the autocorrelation which is still present due to the forcing-related component may affect the results of the test. But since the graphical investigation did not reveal apparent departures from the normal distribution for all sequences (see Figure A3 in Appendix), we refrained from performing the formal tests for the δ-depending sequences.

Another important aspect to discuss is the estimation of the internal

variability of the time-aggregated *single forcing* climate models. Due to the absence of replicates of them, neither $\sigma^2_{\delta_{\text{volcanic}}}$ nor $\sigma^2_{\delta_{\text{land use}}}$ can be estimated directly according to (2.3.3). But for three statistical models the knowledge about $\sigma^2_{\delta_f}$ is essential for the estimation of $\lambda_{11}/\lambda_{21}$. What we can do in this situation is to use the unforced climate model, as it was done, for example, in SUN12. Unfortunately, the estimate of $\widehat{\sigma}^2_{\delta_{\text{unforced}}}$, 0.0143, based on three decadal replicates of x_{unforced}, turned out to be too high for getting a proper solution of the ME- and `j.i.FA(2,1)`-models for all replicates of $x_{\text{land use}}$ and for some replicates of x_{volcanic}. This, of course, impairs the usefulness of our analysis, because the most simple, and therefore the most interesting models, are to be excluded from the numerical experiment. To circumvent this difficulty, the internal variabilities of the x_{volcanic}- and $x_{\text{land use}}$-models were set to the internal variability of x_{E1}, 0.0113, which is motivated under the assumption of an equal internal, i.e. unforced, variability of all climate models involved. Our choice to set $\sigma^2_{\delta_f}$ to the internal variability of x_{E1} instead of the internal variability of x_{E2} can be justified by the fact that $\widehat{\sigma}^2_{\delta_{\text{E1}}} = 0.0113$ and $\widehat{\sigma}^2_{\delta_{\text{E2}}} = 0.0134$ turned out not to differ significantly.

The a priori knowledge about $\sigma^2_{\delta_f}$ is also required by the more complicated `j.i.FA(2,2)`-model. On the other hand, this knowledge is not necessary for the `FA(3,2)`-model, under which the parameter $\sigma^2_{\delta_f}$ can be treated either as free or known. In the former case, the two remaining specific variances, $\sigma^2_{\eta_{\text{internal}}}$ and $\sigma^2_{\delta_{\text{total}}}$, should be specified in advance. Is the latter requirement feasible in our pseudo-proxy study? The answer is yes. Since both $\sigma^2_{\eta_{\text{internal}}}$ and $\sigma^2_{\delta_{\text{total}}}$ are represented by either $\sigma^2_{\delta_{\text{E1}}}$ and/or $\sigma^2_{\delta_{\text{E2}}}$ (for an example see (3.2.1)), they can be directly estimated from the replicates of x_{E1} respective x_{E2}. If in addition $\sigma^2_{\delta_f}$ is treated as known, the model becomes overidentified with one degree of freedom. As a matter of fact, several plausible three-indicator models can be formulated on the basis of model (2.5.16). In Table 1, an overview of the statistical models, analyzed in the present work, is given. In the same table one can find the associated vectors of known hypothesized parameters, i.e. parameters that are not a part of identifiability conditions, but a part of hypotheses, and the associated confidence regions for $\lambda_{11}/\lambda_{21}$. Regarding the overidentified three-indicator models, a model with the best overall fit, provided a solution is proper and interpretable, will be chosen as a final model. As a further remark on Table 1, we specify the known specific variances required for the identifiability:

- $\sigma^2_{\delta_f}$ in the `j.i.ME`- and `j.i.FA(2,1)`-models,
- $\sigma^2_{\delta_f}$, $\sigma^2_{\eta_{\text{internal}}}$ in the `j.i.FA(2,2)`- and `o.i.FA(2,1)`-models,
- $\sigma^2_{\eta_{\text{internal}}}$, $\sigma^2_{\delta_{\text{total}}}$ in all three-indicator models.

`Table 1`. *Overview of the statistical models with the associated vectors of known parameters and the associated confidence regions for* $\lambda_{11}/\lambda_{21}$. *Each model was fitted to data satisfying* $\sigma^2_\epsilon = 0$

Model	Reference number	*df*	Known hypothesized parameters	$CR_{\lambda_{11}/\lambda_{21}}$ is calculated according to
`j.i.ME`	(2.3.1)	0	-	(2.3.12) i.e. the Wald CI
`j.i.FA(2,1)`	(2.3.14)	0	-	(2.3.25) i.e. the Fieller CR
`j.i. FA(2,2)`	(2.5.2)	0	-	(2.3.25)
`o.i.ME`	(2.5.9)	1	$\lambda_{22} = 0$	(2.3.12)
`o.i.FA(2,1)`	(2.5.4)	1	$\lambda_{22} = 0$	(2.3.25)
`j.i.FA(3,2)`	(2.5.16)	0	-	——"——
1st `o.i.FA(3,2)`	——"——	1	$\sigma^2_{\delta_f} = 0.0113$	——"——
2nd `o.i.FA(3,2)`	——"——	1	$\lambda_{22} = 0$	——"——
3d `o.i.FA(3,2)`	——"——	1	$\lambda_{32} = 0$	——"——
4th `o.i.FA(3,2)`	——"——	2	$\sigma^2_{\delta_f} = 0.0113$, $\lambda_{22} = 0$	——"——
5th `o.i.FA(3,2)`	——"——	2	$\sigma^2_{\delta_f} = 0.0113$, $\lambda_{32} = 0$	——"——
1st `o.i.FA(3,1)`	——"——	2	$\lambda_{22} = \lambda_{32} = 0$	——"——
2nd `o.i.FA(3,1)`	——"——	3	$\sigma^2_{\delta_f} = 0.0113$, $\lambda_{22} = \lambda_{32} = 0$	——"——

3.3 Numerical results for data with zero proxy noise

We start our comparative analysis by investigating the performance of the estimators in 'true' conditions, i.e. when 'true' τ is available. All models, presented in Table 1, were fitted to the data sets with zero proxy noise, i.e. $\sigma_\epsilon^2 = 0$. The numerical results are summarized in Summary 1 (see Appendix). As follows from Summary 1, all overidentified models, accepted as final models, have a very good overall fit to data both statistically and heuristically.

To simplify the discussion about the estimates of $\lambda_{11}/\lambda_{21}$, they all are summarized graphically in Figure 2 and 3 together with the observed values of the U_R-statistic. In both figures the estimates obtained under the one-factor models, i.e. the `j.i.ME`-model and the `j.i.FA(2,1)`-model, both associated with the same estimator of $\lambda_{11}/\lambda_{21}$, are separated from the estimates obtained under the two-factor models.

Figure 2 shows the result for the climate model driven by the volcanic forcing, while Figure 3 for $x_{\mathsf{land\text{-}use}}$. Note the different scales of the horizontal axis in the figures.

The number of estimates obtained under the one-factor models is not high. Due to the limited number of replicates of x_{E1} (5 repl.) and x_{E2} (3 repl.), used as pseudo-τ, only eight estimates for each single forcing climate model, x_{volcanic} and $x_{\mathsf{land\text{-}use}}$, are available. Clearly it is not possible to draw definite conclusions based on few estimates, but it is still possible to get some general idea about the performance of the estimators.

For two-factor models, inclusion of the third indicator, x_{total}, led to the larger number of estimates, namely 33 for x_{volcanic} and 30 for $x_{\mathsf{land\text{-}use}}$. For the latter climate model, fitting the three-indicator models to three data sets resulted in inadmissable solutions, why we have 30 estimates instead of 33. It should be remarked, that all members of both ensembles were arranged into pairs randomly in such a way that each pair was associated only with one data set. By doing so, we remove from the analysis data sets containing essentially identical information, and therefore expected to lead to correlated estimates.

Comparing Figure 2 and Figure 3, we first of all note that the performance of all estimators is fairly good when the test statistic, U_R, take on significant values, or equivalently when the covariance between the simulated temperature and the pseudo-τ is significantly different from zero. On the contrary, when the test statistic is not significant, each estimator seems to fail to provide reasonable estimates of $\lambda_{11}/\lambda_{21}$ (at least for one data set).

Theoretically, given that $\lambda_{11}/\lambda_{21} = 1$, the highly significant values of the

U_R-statistic should be caused by the significant variability of the latent factor ξ_{f}, dominating over the noise variability, whereas the insignificant values imply that $H_0 : \sigma^2_{\xi_{\mathrm{f}}} = 0$ is not rejected. This theoretical result is fully confirmed by the numerical results. As follows from Summary 1.1-1.5, where the results for the climate model driven by the volcanic forcing are reported, the leading coefficient a in Eq. 2.31 is positive for all data sets analyzed, which amounts to saying that $\widehat{\sigma}^2_{\xi_{\mathrm{volcanic}}}$ is significantly different from zero (here, at the 5% level). Regarding the climate model driven by the land-use forcing, the result is opposite: the coefficient is negative for all data sets analyzed (see Summary 1.6-1.10), which means that the hypothesis $\sigma^2_{\xi_{\mathrm{f}}} = 0$ is not rejected at the same significance level. The latter fact explains not only the insignificance of the test statistic, but also the unstable estimation of $\lambda_{11}/\lambda_{21}$.

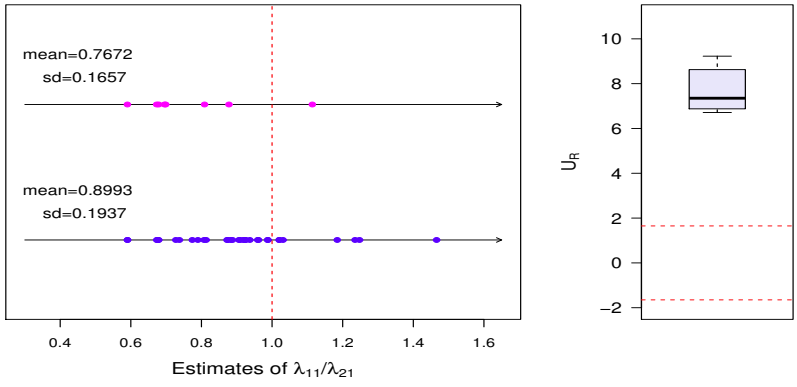

Figure 2. *To the left*: Graphical overview of the estimates of $\lambda_{11}/\lambda_{21}$, obtained under the one-factor models (8 points in magenta) and under various two-factor models (33 points in blue), for the climate model driven by the *volcanic* forcing, x_{vol} (see Summary 1.1-1.5 in Appendix). The dotted vertical line denotes the correct value of $\lambda_{11}/\lambda_{21}$ that is 1. *To the right*: Boxplots for the associated unweighted U_R-statistic (see Appendix). The dotted lines denote the 5% one-sided confidence limits. Calculations are based on the data with zero proxy noise, i.e. $\sigma^2_{\epsilon} = 0$.

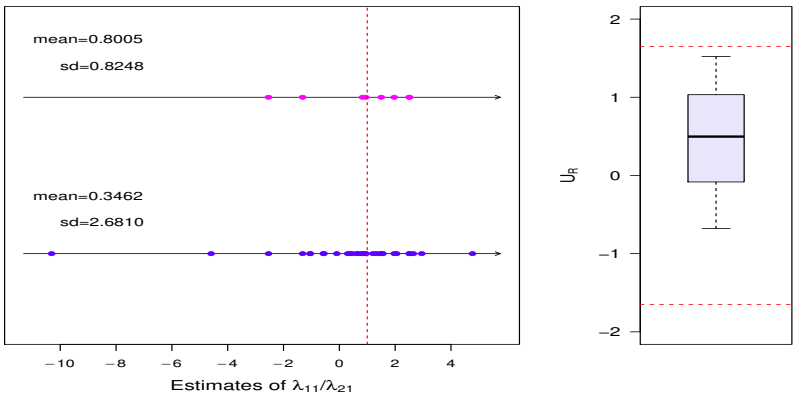

Figure 3. *To the left*: Graphical overview of the estimates of $\lambda_{11}/\lambda_{21}$, obtained under the one-factor models (8 points in magenta) and under various two-factor models (29 points in blue), for the climate model driven by the *land use* forcing, $x_{\text{land-use}}$ (see Summary 1.6-1.10 in Appendix). The dotted vertical line denotes the correct value of $\lambda_{11}/\lambda_{21}$ that is 1. *To the right*: Boxplots for the associated unweighted U_R-statistic (see Appendix). The dotted lines denote the 5% one-sided confidence limits. Calculations are based on the data with zero proxy noise, i.e. $\sigma_\epsilon^2 = 0$.

Theoretically, given that $\lambda_{11}/\lambda_{21} = 1$, the highly significant values of the U_R-statistic should be caused by the significant variability of the latent factor ξ_{f}, dominating over the noise variability, whereas the insignificant values imply that $H_0 : \sigma^2_{\xi_{\text{f}}} = 0$ is not rejected. This theoretical result is fully confirmed by the numerical results. As follows from Summary 1.1-1.5, where the results for the climate model driven by the volcanic forcing are reported, the leading coefficient a in Eq. 2.31 is positive for all data sets analyzed, which amounts to saying that $\widehat{\sigma}^2_{\xi_{\text{volcanic}}}$ is significantly different from zero (here, at the 5% level). Regarding the climate model driven by the land-use forcing, the result is opposite: the coefficient is negative for all data sets analyzed (see Summary 1.6-1.10), which means that the hypothesis $\sigma^2_{\xi_{\text{f}}} = 0$ is not rejected at the same significance level. The latter fact explains not only the insignificance of the test statistic, but also the unstable estimation of $\lambda_{11}/\lambda_{21}$.

Now, let us take a closer look at the estimates associated with the significant U_R-statistic. As indicated by Figure 2, the two-factor models (the

points in blue) on the whole appear to perform better than the one-factor models in the sense that they are closer to the correct value of the amplitude of a forcing effect in a climate model that is 1. This conjecture is also supported by the sample means (compare 0.7672 to 0.8986). Summary 1.1-1.5 together with Figure 4, where the estimates for each two-factor model are shown, give a strong indication that this result is due to deleting structural relations, linked to the insignificant influence of the second latent factor in the two-factor models. In addition, it was found that freeing the parameter $\sigma^2_{\delta_f}$ in the three-indicator models could lead to a further improvement of the estimates of $\lambda_{11}/\lambda_{21}$ and to a good overall fit of the overidentified models, compared to the same models where this specific variance was treated as known.

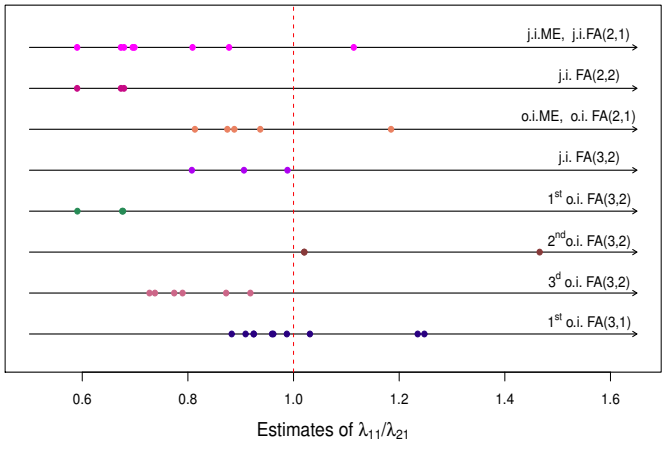

Figure 4. Graphical comparison of the estimates of $\lambda_{11}/\lambda_{21}$, obtained under various statistical models associated with the x_{volcanic}- climate model (see Summary 1.1-1.5 in Appendix). The dotted vertical line denotes the correct value of $\lambda_{11}/\lambda_{21}$ that is 1.

Concomitantly, it was noted that if fitting the `j.i.FA(2,2)`-model to data sets $\{x_{\mathrm{f}\,t}, \tau_t\}$ and $\{x_{\mathrm{f}\,t}, x_{\text{total}\,t}\}$ results in admissable solutions, i.e. all three side conditions are fulfilled, for *both* data sets, then adding x_{total} to $(x_{\mathrm{f}}\,, \tau)$ as a third indicator might lead to an unstable estimation. In many cases, simplifications of such three-indicator models led to a larger underestimation. The instability was especially pronounced when the hypothesis

$H_0 : \sigma^2_{\xi_\mathsf{f}} = 0$ was not rejected, which was observed for the $x_{\mathsf{land\ use}}$-climate model. For this climate model, freeing $\sigma^2_{\delta_\mathsf{f}}$ or setting some loadings to zero could lead to diametrically opposite results, compared to those obtained before modification.

Based on the experience gained during the whole estimation process, it can be recommended to start the estimation process with the estimation of the `j.i.FA(2,2)`-model. If all side conditions are satisfied and the estimate of λ_{22} is statistically significantly different from zero, do not procceed further with the estimation of three-indicator models. If the estimate of λ_{22} is not significant, fit the `o.i.FA(2,1)`-model and assess its overal fit. Only if it is acceptable and if the estimate of λ_{21} is significantly different from zero, a three-indicator model can be fitted, otherwise its estimation stability is questionable. However, bearing in mind that real-world temperature proxies are contaminated with a much larger noise than our pseudo-τ, these recommendations should be taken with care.

Further, based on Figure 4, we may conclude that although the two-factor models seem to perform better than the one-factor models, there is still a tendency to underestimation of the amplitude of a forcing effect. Only the 1^{st}`o.i.FA(3,1)`-model seems to present a reasonable and stable performance among all three-indicator models, indicating that in some situations, adding an additional indicator to a two-indicator model might lead to an improved estimate (compare with the result for the `o.i.FA(2,1)`-model in Figure 4). Regarding the 2^{nd}`o.i.FA(3,2)`-model, the result rather indicates an unstable behaviour, than an improvement.

Having discussed the performance of the estimators, we now compare the performance of the Wald confidence interval and the Fieller confidence region.

First of all, the results in Summary 1.1-1.10 strongly indicate the reliability of the Fieller method, regardless of whether the hypothesis $H_0 : \sigma^2_{\xi_\mathsf{f}} = 0$ is rejected or not. That is, in most cases, we could draw a correct conclusion that the amplitude of a forcing effect in a climate model is of the right size. For the $x_{\mathsf{land\ use}}$-climate model, for which $H_0 : \sigma^2_{\xi_\mathsf{f}} = 0$ was <u>not</u> rejected, most Fieller confidence regions were unbounded/exclusive. Despite of it, the largest part of them were reasonable, more precisely they contain 1 but not 0. Obviously, an unbounded/exclusive confidence region (even reasonable) does not provide precise information about possible values of the amplitude of a forcing effect in a climate model. Nevertheless, the interpretation of a reasonable unbounded/exclusive confidence region is unambiguous. Such a confidence region says us that a particular forcing `f` *is* detected in a climate model simulation, but due to insignificant temperature changes caused by

the forcing, it is not possible to determine with any precision the amplitude of these changes, i.e. the amplitude of its effect.

For a small part of the data sets for which the hypothesis $H_0 : \sigma^2_{\xi_f} = 0$ was not rejected, the Fieller method failed to provide reasonable confidence regions. That is, unbounded/exclusive confidence regions containing both 1 and 0 were observed. According to Summary 1.6-1.10, this failure seems to be associated with three-indicator two-factor models, where the 'true' τ and x_{total} are represented by simulations from different ensembles, i.e. the influence of the second latent factor on these indicators is of different significance, perhaps a common occurrence in real-world data.

Regarding the Wald confidence interval in (2.3.12), we distinguish between a confidence interval for α_f associated with the `j.i.ME`-model and a confidence interval for α_f associated with the `o.i.ME`-model. Since the estimators of α_f under these two models are different, with different asymptotic distributions, we may expect different results concerning such properties as a actual coverage probability of a confidence interval and its expected length.

- `The j.i.FA(2,1)-model`

  For the x_{volcanic}-climate model, for which the hypothesis $H_0 : \sigma^2_{\xi_f} = 0$ was rejected (see Summary 1.1-1.5), the observed confidence intervals do not exhibit any problematic behavior in terms of their lenght, though it seems that the actual coverage probability can be less than 0.95 (more estimates are required to gain more certainty in conclusions).

  When the hypothesis $H_0 : \sigma^2_{\xi_f} = 0$ was not rejected (see the results for the $x_{\mathsf{land\ use}}$-climate model in Summary 1.6-1.10), the method obviously failed to provide reasonable confidence intervals, which illustrates the impact of the Gleser-Hwang effect.

- `The o.i.FA(2,1)-model`

  To begin with, this model had an acceptable overall fit only for the x_{volcanic}-climate model, for which the hypothesis $H_0 : \sigma^2_{\xi_f} = 0$ was rejected. Under this condition, the observed confidence intervals indicate that the performance of the Wald confidence interval is as good as the performance of the Fieller confidence interval.

As mentioned earlier, the properties of the data, analyzed in this section, do not mirror the properties of the real-world proxy z. Therefore, it is not reasonable to draw conclusions about the performance of the statistical models and the methods for constructing confidence regions at this stage of the analysis. What we need to do next is to investigate the impact of adding

a noise to the pseudo-τ on the results observed for the data with the zero proxy noise. The description of this analysis is given in the next section.

3.4 Sensitivity to increasing noise

The question of sensitivity of a statistical model can be addressed by fitting the model to data containing the pseudo-τ distorted by a noise with a considerable variability. Since the results, obtained for the 'true' data, i.e. data with zero proxy noise, clearly show that large residual noise in climate observations, dominating over a *weak* forcing effect, is a great obstacle for precise estimation of the amplitude of such a forcing effect, it is natural to wonder whether increased noise in climate observations impacts the estimation of the amplitude of a *strong* forcing effect in a similar manner. Therefore, in the framework of the present analysis, the sensitivity to increasing noise will be investigated for the statistical models associated with the x_{volcanic}-climate model. Would their estimation still be stable and reliable such that it results in a proper solution, a reasonable confidence interval, and acceptable overall fit when noise is added?

According to Summary 1.1-1.5, the following 'true' statistical models are associated with the x_{volcanic}-climate model:

- One-factor models:
  `j.i.ME`, `j.i.FA(2,1)`
- Two-factor models:
  `j.i.FA(2,2)`, `o.i.ME`, `o.i.FA(2,1)` ,
  `j.i.FA(3,2)`, 1^{st}`o.i.FA(3,2)`, 2^{nd}`o.i.FA(3,2)`, 3^{d}`o.i.FA(3,2)`,
  1^{st}`o.i.FA(3,1)`.

Provided that a large number of data sets is available, the sensitivity of each statistical model above to increasing noise can be investigated by studying:

1. *How close each estimator is to the correct value of* $\lambda_{11}/\lambda_{21}$ *that is 1*, by calculating proportions of admissible/proper estimates satisfying $|\widehat{\lambda}_{11}/\widehat{\lambda}_{21} - 1| \leq s$ for all $s \in (0, 0.9)$ (larger values are not of interest);

2. *Whether the overall fit of the overidentified models remains acceptable*, by calculating the proportion of proper estimates for which the following criteria are satisfied: $GFI \geq 0.9$, $AGFI \geq 0.8$, $SRMR \leq 0.08$, and $CFI \geq 0.95$;

3. *Whether the precision of the estimates of* $\lambda_{11}/\lambda_{21}$ *is acceptable*,
   by calculating the proportion of proper estimates associated with a bounded confidence interval for $\lambda_{11}/\lambda_{21}$, $CI_{\lambda_{11}/\lambda_{21}}$, containing 1 but not 0 (the 5% significance level is considered). Note that observing a bounded confidence interval, constructed according to the Fieller method, corresponds to rejection of the hypothesis H_0 : $\sigma^2_{\xi_\mathsf{f}} = 0$;

4. *Whether the* U_R*-statistic is significantly positive*,
   by calculating the proportion of proper estimates for which the U_R-statistic is statistically significant at the 5% significance level, i.e. the cutoff value is 1.65.

Recall that homoscedasticity is assumed, and all sequences analyzed are decadally resolved.

A single pseudo proxy series $\{z_t\}$ is created by adding noise sequence $\{\epsilon_t\}$ to climate model simulation that represents τ. In this study, the pseudo proxy noise has the characteristics of white noise[15], whose generation is accomplished by taking a sample of n, here $n = 100$, observations on $\epsilon \sim \mathrm{N}(0, \sigma^2_\epsilon)$. To compute the variance of ϵ, σ^2_ϵ, we use a notion of "percent noise by variance", abbr. PNV, which is a common convention in pseudo proxy studies for classifying the level of noise (Smerdon, 2011). It is defined as follows:

$$\mathrm{PNV} = \frac{\sigma^2_\epsilon}{\sigma^2_z} = \frac{\sigma^2_\epsilon}{\sigma^2_\tau + \sigma^2_\epsilon}. \tag{3.4.1}$$

Solving (3.4.1) for σ^2_ϵ, we get

$$\sigma^2_\epsilon = \frac{\mathrm{PNV}}{1 - \mathrm{PNV}} \cdot \sigma^2_\tau. \tag{3.4.2}$$

Realistic values of PNV for real local temperature proxy data that have been used in large-scale temperature reconstructions lie in the range between about 2/3 and 0.94, which means that noise variation accounts for about 67% and 94%, respectively, of the total variation in the observed proxy.

The procedure of creating a pseudo-proxy sequence is repeated N times, here $N \approx 1000$. This approximation arises due to different number of basic data sets. For example, the `j.i.ME`-model is associated with 8 basic data sets with x_{volcanic} as x_f and the pseudo-τ represented by members of different

[15] A sequence of uncorrelated random variables, each with zero mean and variance σ^2 is referred to as **white noise**.

ensembles. It means that 125 $\{\epsilon_t\}$-sequences should be generated for each basic data set and added to the pseudo-τ. At the next step, the statistical model is fitted to new data sets, resulting in 8 samples of estimates of $\lambda_{11}/\lambda_{21}$, each consisting of 125 estimates. By merging the separate samples, we obtain a single sample consisting of exactly $8 \times 125 = 1000$ estimates. Another example is the 3^{nd}`o.i.FA(3,2)`-model that is associated with 6 basic data sets, where the pseudo-τ is represented by only the members of the `E2` ensemble. Generating 167 new data sets on the basis of each basic data set will result in $6 \times 167 = 1002$ estimates of $\lambda_{11}/\lambda_{21}$.

A distinguishing feature of the above approach is that merged samples have a hierarchical structure. Indeed, the variability among the estimates within a merged sample can be attributed to two or three (depending on a statistical model under consideration) sources: the first one is the difference between ensembles, the second one is the difference between climate model simulations within an ensemble, and the third source is the imposed noise, whose variability is completely controlled by us. Obviously, merging estimates into a single sample is justified if and only if their variation due to the first two sources is substantially smaller than their variation due to the imposed noise. Two different situations are exemplified and illustrated in Figure A4 in Appendix by means of the `o.i.FA(2,1)`-model associated with 5 basic data sets, where the pseudo-τ is represented only by the members of the `E1` ensemble. The upper plot of the figure illustrates the situation when the variation among the estimates of $\lambda_{11}/\lambda_{21}$ is mainly due to the difference between replicates within the ensemble, thereby not allowing us to merge the estimates. An opposite situation is shown in the second plot: the variability among the estimates is sufficiently represented by the uncertainty due to the imposed noise, which justifies the merging. Note, the latter result is obtained for one of the PNV values that is typical for high-quality real temperature proxy records, namely $\text{PNV}_z = 2/3$. It turned out that this level of noise is sufficiently high for merging of estimates for each statistical model. Another PNV value of interest is $\text{PNV}_z = 0.94$, for which merging of estimates is also justified. This is a very high noise level, but nevertheless local temperature proxies with that much noise have been regarded by paleoclimatologists as useful. In the following, sensitivity of the statistical models will be investigated for these two noise levels.

The discussion will be based on Figure 5. The plots in the upper row describe the distributions of the estimators in terms of their closeness to the correct value of $\lambda_{11}/\lambda_{21}$ that is 1. More precisely, the proportions of admissible estimates (under each statistical model) satisfying $|\widehat{\lambda}_{11}/\widehat{\lambda}_{21} - 1| \leq s$ are plotted against the deviations s. The higher the proportions, the better, or

equivalently, the smaller s for which the proportions approach 1, the better. The barplots in the second row show the proportion of the admissible estimates for which the last three desired characteristics, listed in the beginning of the section, are true. The boxplots in the third row show the distribution of the U_R-statistic.

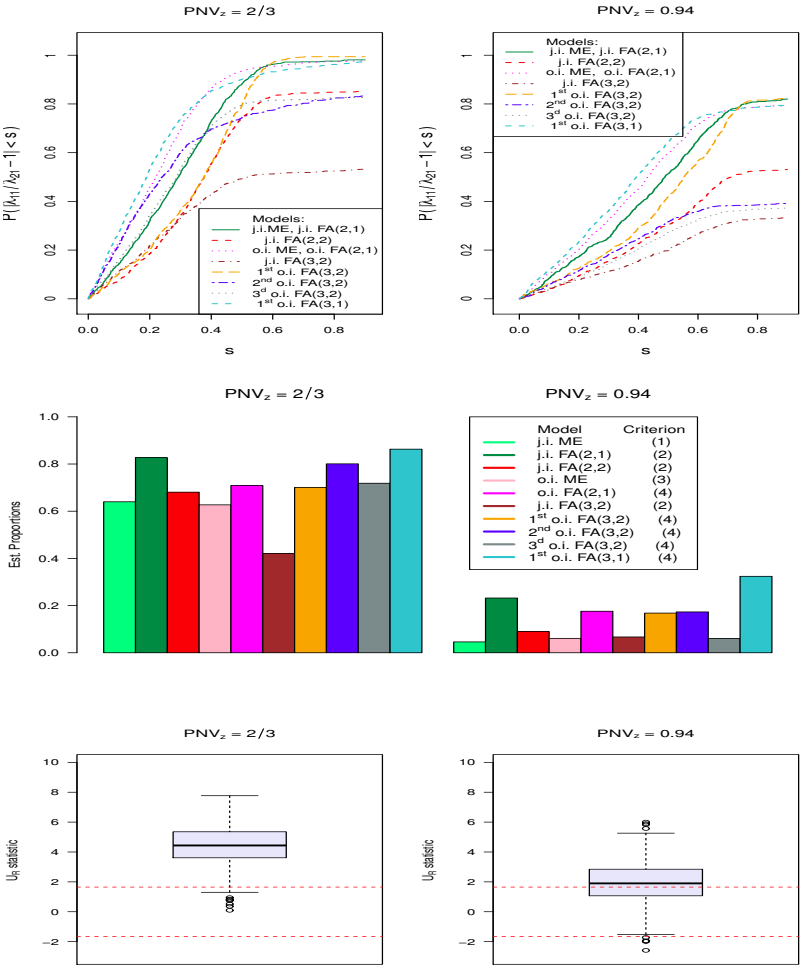

Figure 5. *The upper row:* Empirical probabilities, $P(|\widehat{\lambda}_{11}/\widehat{\lambda}_{21} - 1| \leq s)$ under various statistical models, where $\widehat{\lambda}_{11}/\widehat{\lambda}_{21}$ is a proper estimate. *The second row:* Proportion of

proper estimates of $\lambda_{21}/\lambda_{21}$ for which one of the following criteria is met:

(1) $CI_{\lambda_{11}/\lambda_{21}}$ (2.3.12) contains 1 but not 0, $U_R \geq 1.65$;

(2) $CI_{\lambda_{11}/\lambda_{21}}$ (2.3.24) contains 1 but not 0, $U_R \geq 1.65$;

(3) $GFI \geq 0.9$, $AGFI \geq 0.8$, $SRMR \leq 0.08$, $CFI \geq 0.95$, $CI_{\lambda_{11}/\lambda_{21}}$ (2.3.12) contains 1 but not 0, $U_R \geq 1.65$;

(4) $GFI \geq 0.9$, $AGFI \geq 0.8$, $SRMR \leq 0.08$, $CFI \geq 0.95$, $CI_{\lambda_{11}/\lambda_{21}}$ (2.3.25) contains 1 but not 0, $U_R \geq 1.65$;

Calculations are based on the basic data sets, i.e. data with zero proxy noise, distorted iteratively by noise satisfying either $\text{PNV}_z = 2/3$ or $\text{PNV}_z = 0.94$. Each statistical model is associated with its own set of the basic data sets (see Summary 1). *Remark:* side conditions for a proper solution were explicitly formulated in Sec. 2 for all models except for the three-indicator models. For the latter ones, the completely standardized solutions were checked except singularity cases when no any solution could be obtained. *The third row:* Boxplots for U_R-statistic, describing the correlation between x_{volcanic} and 1,000 pseudo proxy z, constructed by adding $\{\epsilon_t\}_{\mathsf{iteration}_j}$-sequences, $j = 1, 2, \ldots, 125$, to each replicate of x_{E1} and x_{E2}. The 5% one-sided significance levels are shown with dashed lines.

- **$\text{PNV}_z = 2/3$**

According to Figure 5 (both plots in the left panels), the worst performance, or equivalently, the highest sensitivity to the added noise in the pseudo-proxy data, is demonstrated by the `j.i.FA(3,2)`-model. The reasons behind this behavior turned out to be either a singular information matrix, or inadmissible solutions containing a negative estimate of $\sigma^2_{\delta_{\mathsf{f}}}$, i.e. so called Heywood cases. Singularity indicates that a model is underidentified, which amounts to saying that some parameters are not needed. That is, the model should be respecified. Respecification is needed even in the Heywood cases through setting $\sigma^2_{\delta_{\mathsf{f}}}$ to zero. Applied to real-world analysis, it means that if the underlying model is the `j.i.FA(3,2)`-model, it is likely that a substantial amount of noise in the proxy will mask the true relationship between the latent factors and the true temperature, entailing a formulation of another factor model.

Three other three-indicator models, `1st o.i.FA(3,2)`, `2nd o.i.FA(3,2)`, and `3d o.i.FA(3,2)`, exhibit much better performance, compared to the `j.i.FA(3,2)`-model, although singularity and/or Heywood cases still occur. The best performance among the three-indicator models is shown by the `1st o.i.FA(3,1)`-model. To begin with, all 1,000 solutions turned out to be proper, that is, no one singularity case was observed. Further, the associated estimator of $\lambda_{11}/\lambda_{21}$ shows a very good performance in terms of deviations from the correct value of the parameter almost for the whole range of deviations. The estimated probability of observing an acceptable overall fit

and a reasonable bounded confidence interval when the correlation between simulated and observed temperatures is significantly different from zero is the highest, ≈ 0.9, albeit not considerably higher than the corresponding probabilities for the first three models, especially for the 2^{nd}`o.i.FA(3,2)`-model. What it says us is that the increase in the proxy noise affects first of all the overall fit of the 1^{st}`o.i.FA(3,1)`-model.

Singularity was also common for the `j.i.FA(2,2)`-model, though to a less extent than for the `j.i.FA(3,2)`-model. A detailed analysis revealed that in all singularity cases, the third side condition, ensuring a positive estimate of the variance of the second latent factor, i.e. $\widehat{\lambda}_{22} > 0$, was not satisfied. From this, we may conclude that the influence of the increased noise is reflected in the decreased significance of the second latent factor. Once again, we are observing distortion of the true relationships between the variables.

A special attention is drawn to the two groups of models:

Group 1: the `j.i.ME`-model versus the `j.i.FA(2,1)`-model, and
Group 2: the `o.i.ME`-model versus the `o.i.FA(2,1)`-model.

The models within each group have the same estimator of $\lambda_{11}/\lambda_{21}$, but are associated with different methods of constructing a confidence region for the parameter: the ME models with the Wald confidence interval given in (2.3.12), while FA(2,1)-models with the Fieller confidence region obtained by solving inequality (2.3.25).

It turned out that the side conditions for a proper solution associated with the models in each group were fulfilled for each iteration step. Regarding the deviations of the estimates from the correct value of $\lambda_{11}/\lambda_{21}$, the overidentified models demonstrate a better performance than the just-identified models almost for the whole range of the deviations analyzed. In other words, the estimator, associated with the overidentified models in Group 2, is closer to 1. However, the estimated probabilities of observing the same characteristics that the models had for the data with zero proxy noise seem to favor slightly the just-identified models, which indicates the sensitivity of the overall fit of the overidentified models to rising levels of noise. Lastly, comparing the models within each group, we may draw conclusions about the two methods of constructing a confidence interval for $\lambda_{11}/\lambda_{21}$. The observed result speaks in favor of the Fieller method, especially within the first group.

- **$\text{PNV}_z = 0.94$**

Both plots in the right panels in Figure 5 show that the impact of the noise, accounting for 94% of the total variation in the proxy, is huge. All

criteria reflect a strong deterioration in the performance of all statistical models, compared to the results for $\mathrm{PNV}_z = 2/3$: deviations of the estimates from the correct value of the ratio become larger, the number of improper solutions increases, the overall fit of the overidentified models is more often judged as inadequate and unacceptable. Nevertheless, three models are still associated with admissible solutions only. They are: the `j.i.FA(2,1)`-model, the `o.i.FA(2,1)`-model and the `o.i.FA(3,1)`-model. The estimators of the last two models seem to deviate from 1 less than the `j.i.FA(2,1)`-model, although the difference is not substantial. In addition, the estimated probabilities of observing all desired characteristics simultaneously do not differ substantially either, especially between the `j.i.FA(2,1)`- and the `o.i.FA(3,1)`-models. In other words, these two models seem to be equally sensitive to increasing noise.

Further, comparing the results for the models within Group 1 and Group 2 shows clearly the advantage of the Fieller confidence set over the Wald confidence interval. At this point, we recall that the Fieller method, as opposed to the Wald confidence interval which is always bounded, may generate three types of confidence regions, depending on whether the hypothesis $H_0 : \sigma^2_{\xi_\mathrm{f}} = 0$ is rejected or not. Our criterion of an acceptable precision of an estimate is based on the assumption that $H_0 : \sigma^2_{\xi_\mathrm{f}} = 0$ is rejected. However, the distribution of the U_R-statistic for $\mathrm{PNV}_z = 0.94$ in Figure 5 gives rise to a conjecture that the noise with such a large variation affects in decreasing manner not only the significance of the variability of the second latent factor, but also the significance of the variability of the first latent factor. A detailed analysis confirmed this conjecture. That is, rejection of the hypothesis $H_0 : \sigma^2_{\xi_\mathrm{f}} = 0$ is not unlikely for $\mathrm{PNV}_z = 0.94$. Notice that for $\mathrm{PNV}_z = 2/3$, this event was extremely rare.

As known, the Fieller confidence set cannot be bounded when the hypothesis $H_0 : \sigma^2_{\xi_\mathrm{f}} = 0$ is not rejected. But it can generate unbounded/exclusive confidence regions. As observed for the data with, such confidence regions may include 1 but not 0. Without providing precise information about the possible values of the amplitude of a forcing effect in a climate model, they are still reasonable and interpretable from the climatological point of view. Therefore, in order to investigate the performance of the Fieller method comprehensively, the insignificance of the first latent factor, caused by the large proxy noise, should be taken into account. Before performing recalculations, we redefine the criterion for an acceptable precision of a proper estimate associated with a factor model as follows: if $H_0 : \sigma^2_{\xi_\mathrm{f}} = 0$ is rejected, the criterion remains the same, i.e. it is a bounded confidence interval, containing 1 but not 0, but if $H_0 : \sigma^2_{\xi_\mathrm{f}} = 0$ is accepted, a relevant

criterion is an unbounded/exclusive confidence region, still containing 1 but not 0. The result of recalculations are shown in Figure 6. As follows from the figure, the superiority of the Fieller confidence set over the Wald confidence interval becomes even more visible: it is more likely to draw a correct conclusion (either precise or imprecise) about the amplitude of a forcing effect in a climate model when the Fieller method is applied. Furthermore, we also see that it is equally likely as under the `j.i.FA(2,1)`-model as under the `o.i.FA(3,1)`-model, which actually supports the earlier uttered conjecture about an equivalent performance of these two models in the presence of large non-climatic noise.

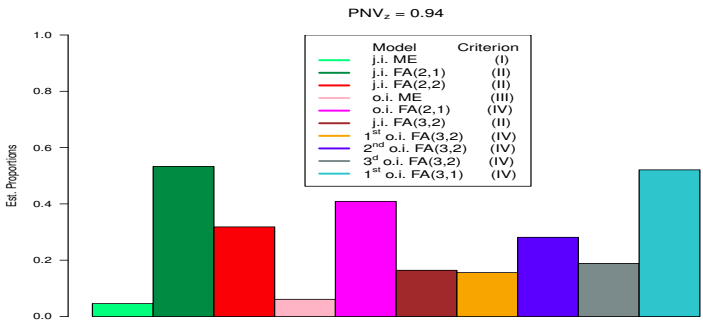

Figure 6. Proportion of estimates of $\lambda_{21}/\lambda_{21}$ under various factor models for which one of the following criteria is met:

`(I)` Solution is proper, $CI_{\lambda_{11}/\lambda_{21}}$ (2.3.12) contains 1 but not 0, $U_R \geq 1.65$.

`(II)` Solution is proper, $CR_{\lambda_{11}/\lambda_{21}}$ (2.3.25) contains 1 but not 0, $U_R \geq 1.65$.

`(III)` Solution is proper, $GFI \geq 0.9$, $AGFI \geq 0.8$, $SRMR \leq 0.08$, $CFI \geq 0.95$, $CI_{\lambda_{11}/\lambda_{21}}$ (2.3.12) contains 1 but not 0, $U_R \geq 1.65$;

`(IV)` Solution is proper, $GFI \geq 0.9$, $AGFI \geq 0.8$, $SRMR \leq 0.08$, $CFI \geq 0.95$, $CR_{\lambda_{11}/\lambda_{21}}$ (2.3.25) contains 1 but not 0, $U_R \geq 1.65$.

Calculations are based on the basic data sets, distorted iteratively by noise satisfying PNV$_z$ = 0.94. Each statistical model is associated with its own set of the basic data sets (see Summary 1). *Remark: the side conditions for a proper solution were explicitly formulated in Sec. 2 for all models except for the three-indicator models. For the latter models, the completely standardized solutions were checked except singularity cases when no any solution could be obtained.*

3.5 Estimation of parameters under heteroscedasticity

Our goal in this section is to test the estimation method that we suggest to apply in the presence of heteroscedasticity. The method is described in Sec. 2. and it consists in replacing the ordinary covariance matrix of indicators by the weighted one.

The pseudo proxy series $\{z_t\}$ with time-varying precision,

$$\sigma_\nu^2(t) = \sigma_\eta^2 + \sigma_\epsilon^2(t) = \sigma_{\xi_{\mathsf{total}\perp\mathsf{f}}}^2 + \sigma_{\eta_{\mathsf{internal}}}^2 + \sigma_\epsilon^2(t), \tag{3.5.1}$$

are created by adding white noise $\{\epsilon_t\}$ to the pseudo-τ such that the variance of ϵ_t accounts for 94% of the total variability in z_t for the period 850-1349 AD, while for the remaining period 1350-1849 AD it accounts for 67%, i.e. $\mathrm{PNV}_z = 2/3$. The choice of basic data sets depends on a statistical model under consideration.

The method was tested on the `j.i.FA(2,1)`-model. Although the model is associated with 8 basic data sets involving the replicates of both ensembles as the pseudo-τ, we, for the sake of simplicity, used only data sets associated with the `E1` ensemble. Once again, the role of the x_{f}-sequence was assigned to the x_{volcanic}-sequence. The vector of free parameters under the `j.i.FA(2,1)`-model in the presence of heteroscedasticity consists of α_{f}, κ and $\sigma_\nu^{2(w)}$, where $\sigma_\nu^{2(w)}$ is the weighted average variability of the proxy z, defined in Sec. 2.4 as follows:

$$\sigma_\nu^{2(w)} = \frac{\sum_{t=1}^n w_t^2 \sigma_\nu^2(t)}{\sum_{t=1}^n w_t^2}, \tag{3.5.2}$$

where the weights w_t are defined in (2.4.12), i.e. $w_t = \sigma_{\eta_{\mathsf{internal}}}^2 \Big/ \sigma_\gamma^2(t)$. Recall, however, the results for the data with zero proxy variance. It was found that the variability of the second latent factor was insignificant for data with τ represented by the `E1` members, that is $\sigma_{\xi_{\mathsf{total}\perp\mathsf{f}}}^2$ in (3.5.1) was estimated as zero. For our pseudo-proxy experiment, this implies that

$$\widehat{\sigma}_\nu^{2(w)} \approx \sigma_\gamma^{2(w)} = \frac{\sum_{t=1}^n w_t^2 \sigma_\gamma^2(t)}{\sum_{t=1}^n w_t^2} = \frac{\sum_{t=1}^{100} w_t^2 \cdot (\sigma_{\eta_{\mathsf{internal}}}^2 + \sigma_\epsilon^2(t))}{\sum_{t=1}^n w_t^2}, \tag{3.5.3}$$

where $\sigma_{\eta_{\mathsf{internal}}}^2$ is independently estimated according to (2.3.3), and the values of $\sigma_\epsilon^2(t)$ are determined according to (3.4.2).

The result above gives us an opportunity to assess the effect of taking heteroscedasticity into account by comparison with the results obtained without assuming heteroscedasticity, more precisely, when the model is fitted to the ordinary sample covariance matrix of the indicators in spite of

the presence of heteroscedasticity. Note, the U_R-statistic under the assumption of homoscedasticity is calculated accordingly, i.e. without applying the weights.

The results of fitting the `j.i.FA(2,1)`-model to heteroscedastic data under the assumption of heteroscedasticity and without assuming it, are summarized in Figure 7. Notice that all estimates of $\lambda_{11}/\lambda_{21}$, obtained under each assumption, were proper. Nevertheless, judging from Figure 7, the effect of taking heteroscedasticity into account is highly positive: (1) the estimates of $\lambda_{11}/\lambda_{21}$ are closer to the correct value of 1, which indicates a higher precision, (2) the estimated probability of observing proper estimates with an acceptable precision, i.e. proper estimates for which the Fieller bounded confidence set contains 1 but not 0, when the correlation between simulated temperatures and the proxy is significantly positive, is almost twice as large as the corresponding probability obtained without taking heteroscedasticity into account, (3) the estimates $\widehat{\sigma}_{\nu}^{2(w)}$:s, obtained without taking heteroscedasticity into account, are too high, compared to the five 'true' values of $\sigma_{\gamma}^{2(w)}$, while the estimates of $\sigma_{\nu}^{2(w)}$, obtained under the assumption of heteroscedasticity, exhibit a good agreement with them.

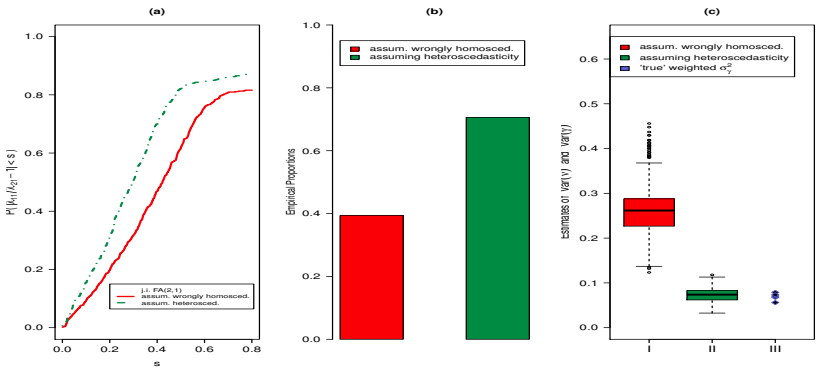

Figure 7. The results of fitting the `j.i.FA(2,1)`-model to heteroscedastic data assuming and without assuming heteroscedasticity:

(a) Empirical probabilities, $P(|\widehat{\lambda}_{11}/\widehat{\lambda}_{21} - 1| \leq s)$, $s \in (0, 0.9)$;

(b) Proportion of proper estimates of $\lambda_{21}/\lambda_{21}$ for which $CI_{\lambda_{11}/\lambda_{21}}$ (2.3.25) contains 1 but not 0, and $U_R \geq 1.65$.

(c) `I`: Boxplot for $\widehat{\sigma}_{\nu}^{2(w)}$, estimated without assuming heteroscedasticity;
`II`: Boxplot for $\widehat{\sigma}_{\nu}^{2(w)}$, estimated under the assumption of heteroscedasticity;
`III`: Five 'true' values of $\sigma_{\gamma}^{2(w)}$, calculated according to (3.5.3).

Calculations are based on the 5 basic data sets $\{x_{\text{volcanic}}, \tau_{\text{E1, repl. i}}\}$, each of which is

distorted iteratively 200 times by noise satisfying $\mathrm{PNV}_z = 0.94$ for $850 - 1349$ AD and $\mathrm{PNV}_z = 2/3$ for $1350 - 1849$ AD.

4 Conclusions

A starting point for this work was the statistical framework, developed by Sundberg et. al (2012), incorporating both climate model simulations, observed climate variables and proxies. This framework provides a theoretical basis for evaluating climate model simulations by estimating the amplitude of a latent forcing effect embedded in the simulations. The aim of the present work was to suggest appropriate statistical methods that can be employed for the estimation.

To this end, several latent factor models were proposed. The models have different structures, differing in the number of observed and/or unobserved (latent) variables (see Table 1 in Sec. 3.2 for an overview). To evaluate and compare their performance we conducted a pseudo-proxy experiment, in which the true unobservable temperature is replaced by selected climate model simulations. The analysis of the data with the zero proxy noise indicated the advantage of more complicated models with two latent factors over the one-factor models. This is due to the possibility to simplify the structure of the former models by eliminating the second latent factor form the models, which may lead to a more precise estimation of the amplitude of a forcing effect in a climate model. However, concerning the three-indicator models, it was found that such a simplification should preferably involve all indicators, like in the `o.i.FA(3,1)`-model, otherwise the estimation procedure might be unstable. This finding was confirmed by analyzing data with added noise: the higher the noise level, the higher chance to observe improper estimates under various FA(3,2)-models, which requires the models to be respecified accordingly.

In the climatological context, the analysis of data with added noise is of the most importance as such data reflect the properties of real-world proxies contaminated with a large non-climatic noise. Increasing noise to such a level when observations consist almost only of noise affected highly negatively all statistical models analyzed. Nevertheless, the results clearly pointed out on two competing models: the `j.i.FA(2,1)`-model and the `o.i.FA(3,1)`-model. Their performance was not only better than the performance of the others but also more or less equal. Based on this result and keeping in mind that the estimation procedure for overidentified models, as opposed to just-identified models, ought to be accompanied by an additional

procedure of assessing the overall model fit to data, which may be challenging for non-statisticians, we may conclude that the simpler model, i.e. the `j.i.FA(2,1)`-model, is preferred for a real-world analysis. In addition, the model demonstrated a quite good performance even under heteroscedasticity.

Note that the `j.i.FA(2,1)`-model has the same estimator of the amplitude of a forcing effect in a climate model as the `j.i.ME`-model, which however demonstrated much worse performance, and therefore cannot be recommended to be applied in reality. The main reason behind it was the use of different methods of constructing a confidence region for the parameter representing the amplitude of a forcing effect. Hence, another important finding of our analysis concerns two methods of constructing a confidence region.

The first method leads always to a bounded confidence interval, referred to as the Wald confidence interval (see (2.3.12)) whereas the second method, based on the Fieller method of finding the confidence interval of the ratio of two normal means (see (2.3.25)), is able to generate not only a bounded confidence intervall but also two types of unbounded confidence regions. The results shows the superiority of the Fieller confidence region, especially when the correlation between simulated temperatures and the proxy/observed temperature is not significantly different from zero. In such cases, the conclusions about the amplitude of a forcing effect in a climate model could be imprecise but still reasonable and interpretable from the climatological point of view, leading to a more comprehensive description of properties of a climate model under consideration.

The discussion in Sec. 2.3.3 has indicated prospective directions of future research. Closer investigation of the `j.i.FA(2,1)`-model is needed in connection with its possible application in the detection and attribution studies. Another interesting topic is a deeper investigation of the differences in performance of statistical models with and without the discrepancy term in the relation between latent forcing effects, suggested by Tingley et al. (2015).

5 References

Allen, M. R. and Stott, P.A.: *Estimating signal amplitudes in optimal fingerprinting, part I: theory*, Climate Dynamics, **21**, pp. 477-491, doi: 10.1007/s00382-003-0313-9, 2003.

Birch, M. W.: *A note in the maximum likelihhod estimation of a linear structural relationship.* J. Amer. Statist. Assoc., **59**, pp. 1175-1178, 1964.

Bollen, K. A.: *Structural equations with latent variables*, Wiley, 1989.

Boomsma, A.: *Reporting Analyses of Covariance Structures*, Structural Equation Modeling, 7(3), pp. 461-483, 2000.

Braconnot, P., Harrison, S. P., Kageyama, M., Bartlein, P. J., Masson-Delmotte, V., Abe-Quchi, A., Otto-Bliesner, B., and Zhao, Y.: *Evaluation of climate models using palaeoclimatic data*, Nature Climate Change, **2**, pp. 417–424, doi:10.1038/nclimate1456, 2012.

Brewer, S., Guiot, J., and Torre, F.: *Mid-Holocene climate change in Europe: a data-model comparison*, Clim Past, **3**, pp. 499-512, 2007.

Cheng, C.-L. and J. W. van Ness: *Statistical regression with measurement error*, Kendall's Library of Statistics, 1999.

Fox, J.: *Structural equation modeling with the* sem package in R, Structural Equation Modeling, 13(3), pp. 465-486, 2006.

Franz, V.H.: *Ratios: A short guide to confidence limits and proper use*, arXiv:0710.2024v1, University of Giessen, 2007.

Fuller, W. A.: *Measurement Error Models*, Wiley, 1987.

Gleser L.J.: *Confidence intervals for the slope in a linear errors-in-variables regression model*, Advances in Multivariate Statistical Analysis (ed. K. Gupta), pp. 85-109, D. Reidel Publishing Company, Dordrecht, 1987.

Gleser L.J., and Hwang J.T.: *The nonexistence of* $100(1-\alpha)\%$ *confidence sets of finite expected diameter in errors-in-variables and related models*, The Annals of Statistics, **15**, No. 4, pp. 1351-1382, 1987.

Goosse, H., P.Y. Barriat, W. Lefebvre, M.F. Loutre, and V. Zunz: *Introduction to climate dynamics and climate modeling*, 2010. Online textbook available at http://www.climate.be/textbook.

Hegerl, G.C., et al. : *Detection of human influence on a new, valited 1500-year temperature reconstruction*, J. Climate, **20**, pp. 650-666, 2007.

Hegerl, G.C., et al. : *Influence of human and natural forcing on European seasonal temperatures*, Nature geoscience, doi: 10.1038/NGEO1057, 2011.

Hind, A., Moberg, A., and Sundberg, R.: *Statistical framework for evaluation of climate simulations by use of climate proxy data from the last millennium - Part 2: A pseudo-proxy study addressing the amplitude of solar forcing*, Clim Past, **8**, doi: 10.5194/cp-8-1355-2012, 2012.

Hu, L. and Bentler, P.M.: *Fit indices in covariance structure modeling: sensitivity to underparameterized model misspecification*, Psychological Methods, **3**, No. 4, pp. 424-453, 1998.

Hu, L. and Bentler, P.M.: *Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives*, Structural Equation Modeling: A Multidisciplinary Journal, 6:1, pp. 1-55, doi: 10.1080/10705519909540118, 1999.

IPCC, 2013: Summary for Policymakers. In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment report of the Intergovernmental Panel on Climate Change* [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University press, Cambridge, United Kingdom and New York, NY, USA.

Jungclaus, J. H., et al.: *Climate and carbon-cycle variability over the last millennium*, Clim. Past, **6**, pp. 723-737, doi:10.5194/cp-6-723-2010, 2010.

Jöreskog, K.G.: *A general approach to confirmatory maximum likelihood factor analysis.* Psychometrika, **34**, 183-202, 1969.

McGuffie, K. and Henderson-Sellers, A.: *A Climate Modelling Primer*, third edition, Wiley, 2005.

Mitchell, J.F.B., Karoly, D.J., Hegerl, G.C., Zwiers, F.W., Allen, M.R., and Marengo, J., 2001: *Detection of climate change and attribution of causes.* In Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change [Houghton, J.T., Y. Ding, D.J. Griggs, M. Noguer, P.J. van der Linden, X. Dai, K. Maskell, and C.A. Johnson (eds.)]. Cambridge University press, Cambridge, United Kingdom and NewYork, NY, USA, 881pp.

Moberg, L. and Sundberg, R.: *Maximum likelihood estimation of a linear functional relationship when one of the departure variances is known*, Scand J Statist 5: 61-64, 1978.

Moberg, A., Sundberg, R., Grudd, H., and Hind, A.: *Statistical framework for evaluation of climate model simulations by use of climate proxy data from the last millennium-Part 3: Practical considerations, relaxed assumptions, and using tree-ring data to address the amplitude of solar forcing*, Clim Past, **11**, pp. 425-448, doi: 10.5194/cp-11-425-2015, 2015

Mulaik, S. A.: *Foundations of Factor Analysis*, 2nd edition, Chapman&Hall/CRC, 2010.

PAGES2k-PMIP3 group: *Continental-scale temperature variability in PMIP3 simulations and PAGES 2k regional temperature reconstructions over the past millennium*, Clim. Past Discuss., **11**, pp. 2483-2555, doi:10.5194/cpd-11-2483-2015, 2015.

Sharma, S.: *Applied multivariate techniques*, Wiley, 1996.

Smerdon, J.E.: *Climate models as a test bed for climate reconstruction methods: pseudoproxy experiments*, WIREs Clim Change, **3**, Issue 1, pp. 63–77, doi: 10.1002/wcc.149, 2012.

Sundberg, R., Moberg, A., and Hind, A.: *Statistical framework for evaluation of climate model simulations by use of climate proxy data from the last millennium-Part 1: Theory*, Clim Past, **8**, pp. 1399 – 1353, 2012, doi: 10.5194/cp-8-1339-2012.

Texier, D., et al. : *Quantifying the role of biosphere-atmosphere feedbacks in climate change: coupled model simulations for 6000 years BP and comparison with palaeodata for northern Eurasia and northern Africa*, Climate Dynamics, **13**, Issue 12, pp 865-881, 1997.

Tingley, M., et al.: *On discriminating between GSM forcing configurations using bayesian reconstruction of late-holocene temperatures*, Journal of Climate, **28**, pp. 8264-8281, doi: 10.1175/JCLI-D-15-0208.1, 2015.

Victor, D.G., D. Zhou, E.H.M. Ahmed, P.K. Dadhich, J.G.J. Oliver, H-H. Rogner, K. Sheikho, and M. Yamaguchi, 2014: Introductory Chapter. In: *Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* [Edenhofer O., R. Pichs-Madruga, Y. Sokona, E. Farahani, S. Kadner, K. Seyboth, A. Adler, I. Baum, S. Brunner, P. Eickemeier, B. Kriemann, J. Savolainen, S. Schlömer, C. von Stechow, T. Zwickel and J.C. Minx (eds.)]. Cambridge University Press, United Kingdom and New York, NY, USA.

6 Appendix

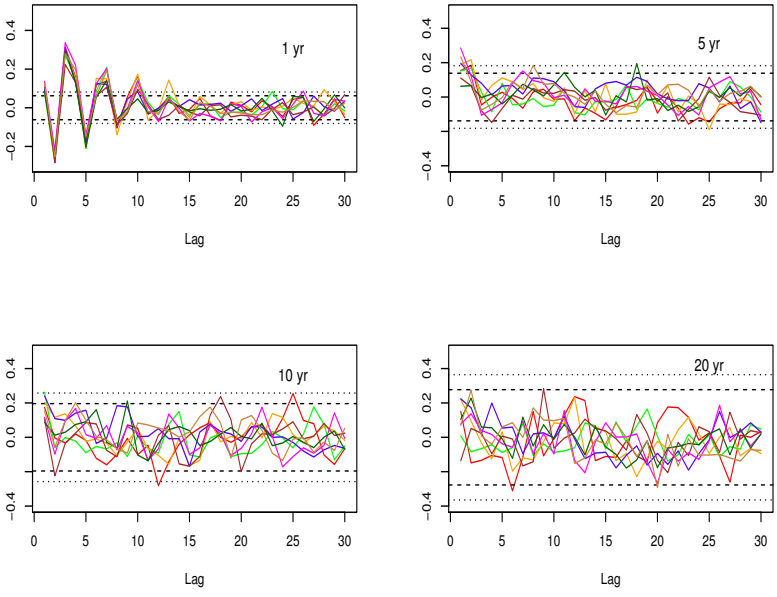

Figure A1. The sample ACF function up to 30 lags for the residual $\{\delta_{\text{Ensemble } t}\} = \{x_{\text{Ensemble, repl.i } t} - \bar{x}_{\text{Ensemble } .t}\}$-sequences (the E1 and E2 ensembles). The two-sided 95% and 99% bounds, denoted by dashed lines, are equal to $\pm 1.96/\sqrt{n}$ and $\pm 2.58/\sqrt{n}$, respectively, where $n = 1000/m$.

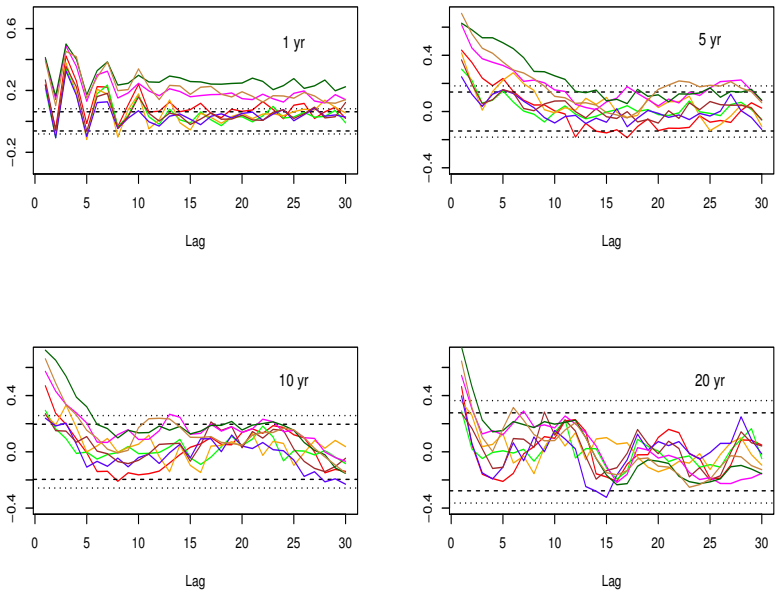

Figure A2. The sample ACF function up to 30 lags for $\{x_{\text{Ensemble, repl.i } t}\}$-sequences (the E1 and E2 ensembles). The two-sided 95% and 99% bounds, denoted by dashed lines, are equal to $\pm 1.96/\sqrt{n}$ and $\pm 2.58/\sqrt{n}$, respectively, where $n = 1000/m$.

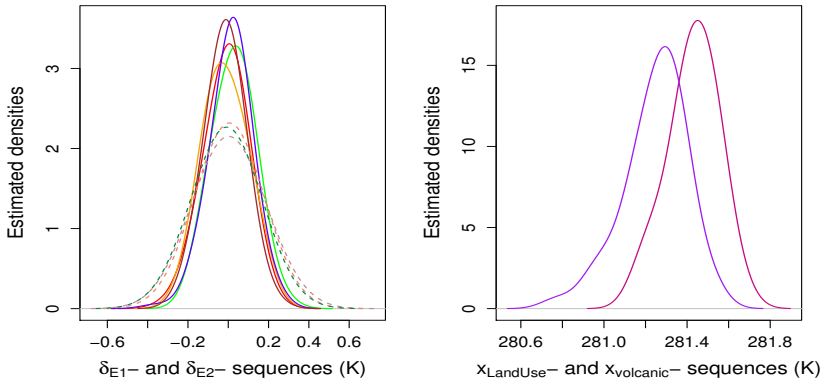

Figure A3. *To the left:* Estimated densities for the $\{\delta_{\text{Ensemble, repli } t}\} \approx \{x_{\text{Ensemble, repl.i } t} - \bar{x}_{\text{Ensemble } .t}\}$- sequences (the E1 and E2 ensembles); *To the right:* Estimated densities for the single forcing simulations. All data have the time unit of 10 years ($m = 10$).

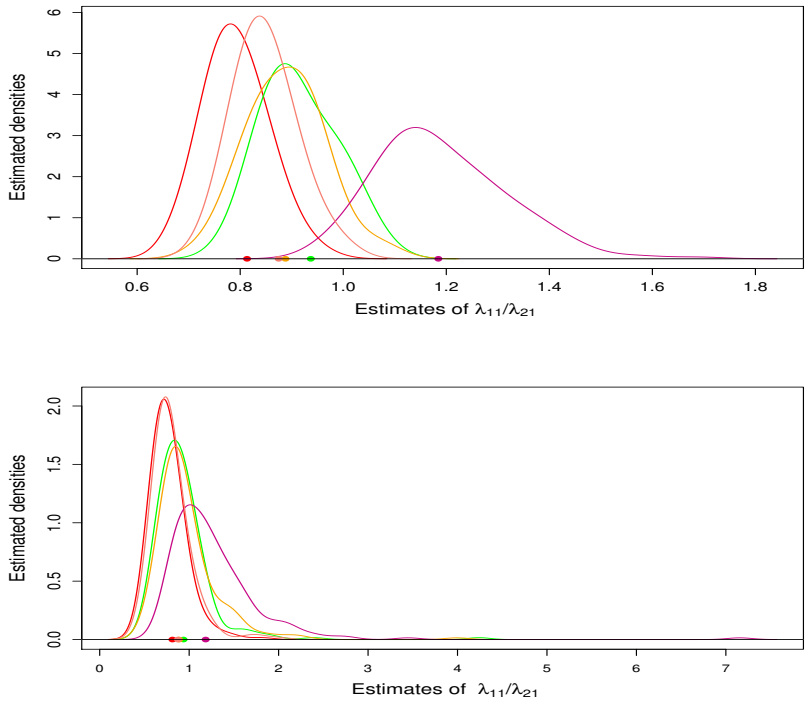

Figure A4. Estimated densities for $\widehat{\lambda}_{11}/\widehat{\lambda}_{21}$ associated with the `o.i.FA(2,1)`-model, and calculated on the basis of data sets obtained by adding noise satisfying $\mathrm{PNV}_z = 0.2$ (the upper plot), and $\mathrm{PNV}_z = 2/3$ (the second plot) to five basic data sets with x_{volcanic} as x_{f} and $x_{\mathsf{E1,\ repl.i}}$ as τ, $i = 1, 2, 3, 4, 5$. The colored dots at the horizontal axis denote $\widehat{\lambda}_{11}/\widehat{\lambda}_{21}$ associated with the basic sets (see Summary 1.1-1.5.).

Summary 1. *Result of fitting the statistical models described in Table 1. Only the final models associated with an admissible/interpretable solution and an acceptable overall fit are presented. The overall fit of the overidentified models was assessed statistically by the* χ^2 *test, involving the* G *test statistic, defined in (2.3.20), and heuristically using four goodness-of-fit indices:* GFI, AGFI, SRMR *and* CFI *(see 2.5.5-2.5.8). The associated* U_R*-statistic was calculated according to (A.1) in Appendix with weights equal to*

1 due to the fact that $\sigma_\epsilon^2 = 0$. *The results are based on the following data sets:*

1.1: (1) $\{x_{\text{volcanic}},\ \tau_{\text{E1, repl.1}}\}$;
(2) $\{x_{\text{volcanic}},\ \tau_{\text{E2, repl.1}}\}$;
(3) $\{x_{\text{volcanic}},\ \tau_{\text{E1, repl.1}},\ x_{\text{E1, repl.j}}\}$, $j = 2, 4$;
(4) $\{x_{\text{volcanic}},\ \tau_{\text{E2, repl.1}},\ x_{\text{E2, repl.j}}\}$, $j = 2, 3$;
(5) $\{x_{\text{volcanic}},\ \tau_{\text{E1, repl.1}},\ x_{\text{E2, repl.j}}\}$, $j = 1, 3$;
(6) $\{x_{\text{volcanic}},\ \tau_{\text{E2, repl.1}},\ x_{\text{E1, repl.j}}\}$, $j = 4, 5$.

$$U_R(x_{\text{volcanic}}, \tau_{\text{E1, repl.1}}) = 8.879, \qquad U_R(x_{\text{volcanic}}, \tau_{\text{E2, repl.1}}) = 6.712$$

Model	Ref. num.	df	Data set	$x_{\text{repl.}j}$	$a > 0$	$\widehat{\lambda}_{11}/\widehat{\lambda}_{21}$	95% CR for $\lambda_{11}/\lambda_{21}$ calculated according to (2.3.12)	(2.3.25)
ME, FA(2,1)	2.3.1&.14	0	(1)	-	TRUE	0.696	(0.441, 0.950)	(0.471, 1.007)
			(2)	-	TRUE	0.673	(0.368, 0.977)	(0.430, 1.122)
o.i.FA(2,1)	2.5.4	1	(1)	-	TRUE	0.813	(0.585, 1.042)	(0.604, 1.077)
j.i.FA(2,2)	2.5.2	0	(2)	-	TRUE	0.673	–	(0.430, 1.122)
1st o.i.FA(3,1)	2.5.16	2	(3)	2	TRUE	0.924	–	(0.714, 1.198)
				4	TRUE	0.883	–	(0.666, 1.161)
1st o.i.FA(3,2)	2.5.16	1	(4)	2	TRUE	0.676	–	(0.431, 1.126)
				3	TRUE	0.677	–	(0.428, 1.143)
j.i.FA(3,2)	2.5.16	0	(5)	1	TRUE	0.989	–	(0.595, 1.638)
				3	TRUE	0.906	–	(0.641, 1.265)
3d o.i.FA(3,2)	2.5.16	1	(6)	4	TRUE	0.918	–	(0.574, 1.592)
				5	TRUE	0.738	–	(0.525, 1.063)

where a is the leading coefficient in inequality (2.3.25).
If $a > 0$ is TRUE, $H_0 : \sigma_{\xi_{\text{f}}}^2 = 0$ is rejected.

Model	Ref. num.	df	Data set	$x_{\text{repl.}j}$	G	p-value	GFI	$AGFI$	$SRMR$	CFI
o.i.FA(2,1)	(2.5.4)	1	(1)	-	1.606	0.205	0.983	0.949	0.039	0.99
1st o.i.FA(3,1)	(2.5.16)	2	(3)	2	0.220	0.896	0.999	0.996	0.010	1
				4	0.117	0.941	0.999	0.998	0.009	1
1st o.i.FA(3,2)	(2.5.16)	1	(4)	2	0.192	0.662	0.999	0.992	0.005	1
				3	1.987	0.159	0.987	0.921	0.021	0.99
3d o.i.FA(3,2)	(2.5.16)	1	(6)	4	0.450	0.502	0.997	0.982	0.020	1
				5	4.596	0.032	0.971	0.829	0.041	0.96

to be continued on the next page

1.2: (1) $\{x_{\text{volcanic}},\ \tau_{\text{E1, repl.2}}\}$,
(2) $\{x_{\text{volcanic}},\ \tau_{\text{E2, repl.2}}\}$,
(3) $\{x_{\text{volcanic}},\ \tau_{\text{E1, repl.2}},\ x_{\text{E1, repl.j}}\}$, $j = 3, 4$;
(4) $\{x_{\text{volcanic}},\ \tau_{\text{E2, repl.2}},\ x_{\text{E2, repl.j}}\}$, $j = 3$;
(5) $\{x_{\text{volcanic}},\ \tau_{\text{E1, repl.2}},\ x_{\text{E2, repl.j}}\}$, $j = 2, 3$;
(6) $\{x_{\text{volcanic}},\ \tau_{\text{E2, repl.2}},\ x_{\text{E1, repl.j}}\}$, $j = 1, 5$.

$$U_R(x_{\text{volcanic}}, \tau_{\text{E1, repl.2}}) = 8.373, \qquad U_R(x_{\text{volcanic}}, \tau_{\text{E2, repl.2}}) = 7.176$$

Model	Ref. num.	df	Data set	$x_{\text{repl.}j}$	$a > 0$	$\hat{\lambda}_{11}/\hat{\lambda}_{21}$	95% CI for $\lambda_{11}/\lambda_{21}$ calculated according to (2.3.12)	(2.3.25)
ME, FA(2,1)	2.3.1&.14	0	(1)	-	TRUE	0.809	(0.500, 1.118)	(0.542, 1.199)
			(2)	-	TRUE	0.590	(0.337, 0.844)	(0.383, 0.945)
o.i.FA(2,1)	2.5.4	1	(1)	-	TRUE	0.937	(0.648, 1.226)	(0.679, 1.285)
j.i.FA(2,2)	2.5.2	0	(2)	-	TRUE	0.590	–	(0.383, 0.945)
1^{st}o.i.FA(3,1)	2.5.16	2	(3)	3	TRUE	0.962	–	(0.706, 1.303)
				4	TRUE	0.959	–	(0.695, 1.312)
1^{st}o.i.FA(3,2)	2.5.16	1	(4)	3	TRUE	0.591	–	(0.382, 0.951)
2^{nd}o.i.FA(3,2)	2.5.16	1	(5)	2	TRUE	1.021	–	(0.746, 1.394)
				3	TRUE	1.020	–	(0.743, 1.399)
3^{d}o.i.FA(3,2)	2.5.16	1	(6)	1	TRUE	0.727	–	(0.519, 1.056)
				5	TRUE	0.790	–	(0.565, 1.162)

where a is the leading coefficient in inequality (2.3.25).
If $a > 0$ is TRUE, $H_0 : \sigma^2_{\xi_{\text{f}}} = 0$ is rejected.

Model	Ref. num.	df	Data set	$x_{\text{repl.}j}$	G	p-value	GFI	$AGFI$	$SRMR$	CFI
o.i.FA(2,1)	2.5.4	1	(1)	-	1.396	0.237	0.985	0.956	0.038	0.99
1^{st}o.i.FA(3,1)	2.5.16	2	(3)	3	0.963	0.618	0.994	0.981	0.021	1
				4	1.860	0.395	0.988	0.964	0.031	1
1^{st}o.i.FA(3,2)	2.5.16	1	(4)	3	4.331	0.037	0.972	0.831	0.022	0.98
2^{nd}o.i.FA(3,2)	2.5.16	1	(5)	2	0.193	0.661	0.999	0.992	0.010	1
				3	0.211	0.646	0.999	0.992	0.010	1
3^{d}o.i.FA(3,2)	2.5.16	1	(6)	1	0.124	0.725	0.999	0.995	0.007	1
				5	0.606	0.436	0.996	0.976	0.016	1

to be continued on the next page

1.3: (1) $\{x_{\text{volcanic}},\ \tau_{\text{E1, repl.3}}\}$;
(2) $\{x_{\text{volcanic}},\ \tau_{\text{E2, repl.3}}\}$;
(3) $\{x_{\text{volcanic}},\ \tau_{\text{E1, repl.3}},\ x_{\text{E1, repl.j}}\}$, $j = 1$;
(4) $\{x_{\text{volcanic}},\ \tau_{\text{E2, repl.3}},\ x_{\text{E1, repl.j}}\}$, $j = 3, 4, 5$.

$$U_R(x_{\text{volcanic}}, \tau_{\text{E1, repl.3}}) = 7.524, \qquad U_R(x_{\text{volcanic}}, \tau_{\text{E2, repl.3}}) = 6.887$$

Model	Ref. num.	df	Data set	$x_{\text{repl.}j}$	$a > 0$	$\hat{\lambda}_{11}/\hat{\lambda}_{21}$	95% CI for $\lambda_{11}/\lambda_{21}$ calculated according to (2.3.12)	(2.3.25)
ME, FA(2,1)	2.3.1&.14	0	(1)	-	TRUE	0.878	(0.515, 1.241)	(0.575, 1.369)
		0	(2)	-	TRUE	0.679	(0.378, 0.981)	(0.436, 1.113)
o.i.FA(2,1)	2.5.4	1	(1)	-	TRUE	0.888	(0.610, 1.166)	(0.638, 1.220)
j.i.FA(2,2)	2.5.2	0	(2)	-	TRUE	0.679	–	(0.436, 1.113)
1st o.i.FA(3,1)	2.5.16	2	(3)	1	TRUE	1.031	–	(0.770, 1.387)
3st o.i.FA(3,2)		1	(4)	3	TRUE	0.774	–	(0.520, 1.191)
				4	TRUE	0.872	–	(0.554, 1.456)
j.i.FA(3,2)		0		5	TRUE	0.808	–	(0.576, 1.174)

where a is the leading coefficient in inequality (2.3.25).
If $a > 0$ is TRUE, $H_0 : \sigma^2_{\xi_f} = 0$ is rejected.

Model	Ref. num.	df	Data set	$x_{\text{repl.}j}$	G	p-value	GFI	$AGFI$	$SRMR$	CFI
o.i.FA(2,1)	2.5.4	1	(1)	-	0.007	0.936	0.999	0.999	0.003	1
1st o.i.FA(3,1)	2.5.16	2	(3)	1	1.329	0.514	0.991	0.973	0.026	1
3d o.i.FA(3,2)		1	(4)	3	0.349	0.555	0.997	0.986	0.013	1
				4	0.171	0.679	0.998	0.993	0.012	1

to be continued on the next page

1.4: (1) $\{x_{\text{volcanic}},\ \tau_{\text{E1, repl.4}}\}$;
(2) $\{x_{\text{volcanic}},\ \tau_{\text{E1, repl.4}},\ x_{\text{E1, repl.j}}\};\ j = 3, 5.$
(3) $\{x_{\text{volcanic}},\ \tau_{\text{E1, repl.4}},\ x_{\text{E2, repl.j}}\}.\ j = 2.$

$$U_R(x_{\text{volcanic}},\ \tau_{\text{E1, repl.4}}) = 6.865$$

Model	Ref. num.	df	Data set	$x_{\text{repl.}j}$	$a > 0$	$\hat{\lambda}_{11}/\hat{\lambda}_{21}$	95% CI for $\lambda_{11}/\lambda_{21}$ calculated according to (2.3.12)	(2.3.25)
ME,FA(2,1)	2.3.1&.14	0	(1)	-	TRUE	1.114	(0.618, 1.610)	(0.715, 1.830)
o.i.FA(2,1)	2.5.4	1	(1)	-	TRUE	1.185	(0.740, 1.630)	(0.811, 1.776)
1^{st}o.i.FA(3,1)	2.5.16	3	(2)	3	TRUE	1.235	–	(0.849, 1.828)
				5	TRUE	1.248	–	(0.900, 1.773)
2^{nd}o.i.FA(3,2)	2.5.16	1	(3)	2	TRUE	1.466	–	(0.975, 2.282)

where a is the leading coefficient in inequality (2.3.25).
If $a > 0$ is TRUE, $H_0 : \sigma^2_{\xi_{\text{f}}} = 0$ is rejected.

Model	Ref. num.	df	Data set	$x_{\text{repl.}j}$	G	p-value	GFI	$AGFI$	$SRMR$	CFI
o.i. FA(2,1)	2.5.4	1	(1)	-	0.207	0.649	0.998	0.994	0.018	1
1^{st}o.i.FA(3,1)	2.5.16	3	(2)	3	0.169	0.919	0.999	0.997	0.009	1
				5	6.646	0.036	0.959	0.878	0.051	0.96
2^{nd}o.i.FA(3,2)	2.5.16	1	(3)	2	0.436	0.509	0.987	0.982	0.020	1

to be continued on the next page

1.5: (1) $\{x_{\text{volcanic}},\ \tau_{\text{E1, repl.5}}\}$,
(2) $\{x_{\text{volcanic}},\ \tau_{\text{E1, repl.5}},\ \tau_{\text{E1, repl.}j}\}$, $j = 1, 2, 3$;

$$U_R(x_{\text{volcanic}}, \tau_{\text{E1, repl.5}}) = 9.226$$

Model	Ref. num.	df	Data set	$x_{\text{repl.}j}$	$a > 0$	$\hat{\lambda}_{11}/\hat{\lambda}_{21}$	95% CI for $\lambda_{11}/\lambda_{21}$ calculated according to (2.3.12)	(2.3.25)
ME, FA(2,1)	2.3.1&.14	0	(1)	-	TRUE	0.699	(0.450, 0.948)	(0.476, 0.997)
o.i.FA(2,1)	2.5.4	1	(1)	-	TRUE	0.875	(0.625, 1.125)	(0.647, 1.168)
1st o.i.FA(3,1)	2.5.16	2	(2)	1	TRUE	0.909	–	(0.702, 1.177)
				2	TRUE	0.987	–	(0.764, 1.285)
				3	TRUE	0.925	–	(0.764, 1.214)

where a is the leading coefficient in inequality (2.3.25).
If $a > 0$ is TRUE, $H_0 : \sigma^2_{\xi_f} = 0$ is rejected.

Model	Ref. num.	Data set	$x_{\text{total, repl.}j}$	G	p-value	GFI	$AGFI$	$SRMR$	CFI
o.i.FA(2,1)	2.5.4	(1)	-	3.780	0.051	0.959	0.878	0.059	0.95
1st o.i.FA(3,1)	2.5.16	(2)	1	3.879	0.144	0.975	0.925	0.037	0.99
			2	0.302	0.860	0.999	0.994	0.011	1
			3	2.297	0.317	0.985	0.955	0.031	0.99

to be continued on the next page

1.6: (1) $\{x_{\text{land-use}}, \tau_{\text{E1, repl.1}}\}$;
(2) $\{x_{\text{land-use}}, \tau_{\text{E2, repl.1}}\}$;
(3) $\{x_{\text{land-use}}, \tau_{\text{E1, repl.1}}, x_{\text{E1, repl.j}}\}$, $j = 2, 4$;
(4) $\{x_{\text{land-use}}, \tau_{\text{E2, repl.1}}, x_{\text{E2, repl.j}}\}$, $j = 2, 3$;
(5) $\{x_{\text{land-use}}, \tau_{\text{E1, repl.1}}, x_{\text{E2, repl.j}}\}$, $j = 1, 3$.
(6) $\{x_{\text{land-use}}, \tau_{\text{E2, repl.1}}, x_{\text{E1, repl.j}}\}$, $j = 4, 5$.

$U_R(x_{\text{land-use}}, \tau_{\text{E1, repl.1}}) = 1.246, \qquad U_R(x_{\text{land-use}}, \tau_{\text{E2, repl.1}}) = 0.3436$

Model	Ref. num.	df	Data set	$x_{\text{repl.}j}$	$a > 0$	$\hat{\lambda}_{11}/\hat{\lambda}_{21}$	95% CI for $\lambda_{11}/\lambda_{21}$ calculated according to (2.3.12)	(2.3.25)
ME, FA(2,1)	2.3.1&.14	0	(1)	-	FALSE	0.946	(-1.16, 3.05)	$\{(-\infty, -1.34),\ (0.08, \infty)\}$
			(2)	-	FALSE	2.511	(-13.4, 18.4)	$\{(-\infty, -0.33),\ (0.13, \infty)\}$
j.i.FA(2,2)	2.5.2	0	(1)	-	FALSE	0.946	–	$\{(-\infty, -1.34),\ (0.07, \infty)\}$
			(2)	-	FALSE	2.511	–	$\{(-\infty, -0.33),\ (0.13, \infty)\}$
1st o.i.FA(3,2)	2.5.16	1	(3)	2	FALSE	1.204	–	$\{(-\infty, -1.35),\ (0.18, \infty)\}$
				4	FALSE	0.948	–	$\{(-\infty, -1.37),\ (0.07, \infty)\}$
			(4)	2	FALSE	2.651	–	$\{(-\infty, -0.32),\ (0.13, \infty)\}$
				3	FALSE	1.561	–	$\{(-\infty, -0.26),\ (0.09, \infty)\}$
			(5)	1	FALSE	0.467	–	$\{(-\infty, -4.8),\ (-0.01, \infty)\}$
				3	FALSE	0.685	–	$\{(-\infty, -2.3),\ (0.11, \infty)\}$
			(6)	4	FALSE	0.300	–	$\{(-\infty,\ \infty)\}$
				5	FALSE	2.952	–	$\{(-\infty,\ \infty)\}$

where a is the leading coefficient in inequality (2.3.25).
If $a > 0$ is FALSE, $H_0 : \sigma^2_{\xi_{\text{f}}} = 0$ is not rejected.

Model	Ref. num.	df	Data set	$x_{\text{repl.}j}$	G	p-value	GFI	$AGFI$	$SRMR$	CFI
1st o.i.FA(3,2)	2.5.16	1	(3)	2	0.267	0.605	0.998	0.989	0.014	1
				4	0.000	0.991	1.000	1.000	0.000	1
			(4)	2	0.235	0.628	0.998	0.991	0.006	1
				3	0.191	0.662	0.99	0.992	0.014	1
			(5)	1	0.525	0.469	0.996	0.979	0.029	1
				3	0.134	0.714	0.999	0.995	0.015	1
			(6)	4	1.400	0.237	0.991	0.951	0.059	1
				5	0.700	0.403	0.995	0.972	0.028	1

to be continued on the next page

1.7: (1) $\{x_{\text{land-use}}, \tau_{\text{E1, repl.2}}\}$;
(2) $\{x_{\text{land-use}}, \tau_{\text{E2, repl.2}}\}$;
(3) $\{x_{\text{land-use}}, \tau_{\text{E1, repl.2}}, x_{\text{E1, repl.j}}\}$, $j = 3, 4$;
(4) $\{x_{\text{land-use}}, \tau_{\text{E2, repl.2}}, x_{\text{E2, repl.j}}\}$, $j = 3$;
(5) $\{x_{\text{land-use}}, \tau_{\text{E1, repl.2}}, x_{\text{E2, repl.j}}\}$, $j = 2, 3$;
(6) $\{x_{\text{land-use}}, \tau_{\text{E2, repl.2}}, x_{\text{E1, repl.j}}\}$, $j = 1, 5$.

$$U_R(x_{\text{land-use}}, \tau_{\text{E1, repl.2}}) = -0.510, \qquad U_R(x_{\text{land-use}}, \tau_{\text{E2, repl.2}}) = 0.411$$

Model	Ref. num.	df	Data set	$x_{\text{repl.}j}$	$a > 0$	$\hat{\lambda}_{11}/\hat{\lambda}_{21}$	95% CI for $\lambda_{11}/\lambda_{21}$ calculated according to (2.3.12)	(2.3.25)
ME, FA(2,1)	2.3.1&.14	0	(1)	-	FALSE	-2.54	(-13.7, 8.6)	$\{(-\infty, -0.15),\ (0.60, \infty)\}$
			(2)	-	FALSE	1.97	(-8.5, 12.5)	$\{(-\infty, -0.33),\ (0.11, \infty)\}$
j.i.FA(2,2)	2.5.2	0	(1)	-	FALSE	-2.54	–	$\{(-\infty, -0.15),\ (0.61, \infty)\}$
			(2)	-	FALSE	1.97	–	$\{(-\infty, -0.33),\ (0.11, \infty)\}$
1st o.i.FA(3,2)	2.5.16	1	(3)	3	FALSE	-10.31	–	$\{(-\infty, -0.49),\ (0.57, \infty)\}$
				4	FALSE	-4.60	–	$\{(-\infty, -0.27),\ (0.52, \infty)\}$
			(5)	2	FALSE	-0.58	–	$\{-\infty,\ \infty\}$
				3	FALSE	-0.09	–	$\{-\infty,\ \infty\}$
			(6)	1	FALSE	0.82	–	$\{-\infty,\ \infty\}$
				5	FALSE	0.37	–	$\{-\infty,\ \infty\}$

where a is the leading coefficient in inequality (2.3.25).
If $a > 0$ is FALSE, $H_0 : \sigma^2_{\xi_f} = 0$ is not rejected.

Model	Ref. num.	df	Data set	$x_{\text{repl.}j}$	G	p-value	GFI	$AGFI$	$SRMR$	CFI
1st o.i.FA(3,2)	2.5.16	1	(3)	3	1.584	0.208	0.990	0.937	0.031	0.99
				4	1.997	0.158	0.987	0.921	0.036	0.97
			(5)	2	0.936	0.333	0.994	0.962	0.043	1
				3	1.498	0.221	0.990	0.940	0.063	0.98
			(6)	1	0.897	0.343	0.994	0.964	0.044	1
				5	1.286	0.257	0.991	0.949	0.056	0.99

to be continued on the next page

1.8: (1) $\{x_{\text{land-use}}, \tau_{\text{E1, repl.3}}\}$;
(2) $\{x_{\text{land-use}}, \tau_{\text{E2, repl.3}}\}$;
(3) $\{x_{\text{land-use}}, \tau_{\text{E1, repl.3}}, x_{\text{E1, repl.j}}\}$, $j = 1$;
(4) $\{x_{\text{land-use}}, \tau_{\text{E2, repl.3}}, x_{\text{E1, repl.j}}\}$, $j = 3, 4, 5$;

$$U_R(x_{\text{land-use}}, \tau_{\text{E1, repl.3}}) = 1.521, \qquad U_R(x_{\text{land-use}}, \tau_{\text{E2, repl.3}}) = -0.679$$

Model	Ref. num.	df	Data set	$x_{\text{repl.}j}$	$a > 0$	$\hat{\lambda}_{11}/\hat{\lambda}_{21}$	95% CI for $\lambda_{11}/\lambda_{21}$ calculated according to (2.3.12)	(2.3.25)
ME, FA(2,1)	2.3.1&.14	0	(1)	-	FALSE	0.83	(-0.83, 2.49)	$\{(-\infty, -2.01),\ (0.07, \infty)\}$
			(2)	-	FALSE	-1.32	(-5.83, 3.20)	$\{(-\infty, -0.09),\ (0.52, \infty)\}$
j.i.FA(2,2)	2.5.2	0	(1)	-	FALSE	0.83	–	$\{(-\infty, -2.01),\ (0.07, \infty)\}$
			(2)	-	FALSE	-1.32	–	$\{(-\infty, -0.09),\ (0.52, \infty)\}$
1st o.i.FA(3,2)	2.5.16	1	(3)	1	FALSE	0.62	–	$\{(-\infty, -3.70),\ (0.02, \infty)\}$
			(4)	3	FALSE	-1.04	–	$\{(-\infty, -0.15),\ (0.46, \infty)\}$
				4	FALSE	-0.55	–	$\{(-\infty, 0.08),\ (0.23, \infty)\}$
				5	FALSE	-1.05	–	$\{(-\infty, -0.10)\ , (0.48, \infty)\}$

where a is the leading coefficient in inequality (2.3.25).
If $a > 0$ is FALSE, $H_0 : \sigma^2_{\xi_f} = 0$ is not rejected.

Model	Ref. num.	df	Data set	$x_{\text{repl.}j}$	G	p-value	GFI	$AGFI$	$SRMR$	CFI
1st o.i.FA(3,2)	2.5.16	1	(3)	1	0.415	0.519	0.997	0.983	0.017	1
			(4)	3	0.075	0.784	0.999	0.997	0.011	1
				4	0.724	0.395	0.996	0.971	0.037	1
				5	0.084	0.772	0.999	0.997	0.010	1

to be continued on the next page

1.9: (1) $\{x_{\text{land-use}},\ \tau_{\text{E1, repl.4}}\}$;
(2) $\{x_{\text{land-use}},\ \tau_{\text{E1, repl.4}},\ x_{\text{E1, repl.j}}\}$, $j = 3, 5$;
(3) $\{x_{\text{land-use}},\ \tau_{\text{E1, repl.4}},\ x_{\text{E2, repl.j}}\}$, $j = 2$.

$$U_R(x_{\text{land-use}}, \tau_{\text{E1, repl.4}}) = 0.583,$$

Model	Ref. num.	df	Data set	$x_{\text{repl.}j}$	$a > 0$	$\widehat{\lambda}_{11}/\widehat{\lambda}_{21}$	95% CI for $\lambda_{11}/\lambda_{21}$ calculated according to (2.3.12)	(2.3.25)
ME, FA(2,1)	2.3.1&.14	0	(1)	-	FALSE	2.50	(-7.3, 12.3)	$\{(-\infty, -0.75),\ (0.16, \infty)\}$
j.i.FA(2,2)	2.5.2	0	(1)	-	FALSE	2.50	–	$\{(-\infty, -0.75),\ (0.16, \infty)\}$
j.i.FA(3,2)	2.5.16	0	(2)	3	FALSE	4.77	–	$\{-\infty,\ \infty\}$

where a is the leading coefficient in inequality (2.3.25).
If $a > 0$ is FALSE, $H_0 : \sigma^2_{\xi_{\text{f}}} = 0$ is not rejected.

1.10: (1) $\{x_{\text{land-use}},\ \tau_{\text{E1, repl.5}}\}$;
(2) $\{x_{\text{land-use}},\ \tau_{\text{E1, repl.5}},\ x_{\text{E1, repl.}j}\}, j = 1, 2, 3.$

$$U_R(x_{\text{land-use}}, \tau_{\text{E1, repl.5}}) = 0.820,$$

Model	Ref. num.	df	Data set	$x_{\text{repl.}j}$	$a > 0$	$\widehat{\lambda}_{11}/\widehat{\lambda}_{21}$	95% CI for $\lambda_{11}/\lambda_{21}$ calculated according to (2.3.12)	(2.3.25)
ME,FA(2,1)	2.3.1&.14	0	(1)	-	FALSE	1.51	(-2.93, 5.93)	$\{(-\infty, -0.85),\ (0.11, \infty)\}$
j.i.FA(2,2)	2.5.2	0	(1)	-	FALSE	1.51	–	$\{(-\infty, -0.85),\ (0.11, \infty)\}$
1^{st}o.i.FA(3,2)	2.5.16	1	(2)	1	FALSE	1.32	–	$\{(-\infty, -0.95),\ (0.10, \infty)\}$
				2	FALSE	2.05	–	$\{(-\infty, -0.79),\ (0.21, \infty)\}$
				3	FALSE	1.46	–	$\{(-\infty, -0.94),\ (0.12, \infty)\}$

where a is the leading coefficient in inequality (2.3.25).
If $a > 0$ is FALSE, $H_0 : \sigma^2_{\xi_{\text{f}}} = 0$ is not rejected.

Model	Ref. num.	df	Data set	$x_{\text{repl.}j}$	G	p-value	GFI	$AGFI$	$SRMR$	CFI
1^{st}o.i.FA(3,2)	2.5.16	1	(2)	1	3.634	0.057	0.976	0.857	0.035	0.96
				2	0.452	0.502	0.997	0.982	0.016	1
				3	0.031	0.860	0.999	0.999	0.004	1

The correlation test-statistic U_R

This statistic is of regression type and allows to test the null hypothesis whether the climate model under consideration does not explain any of the temporal variation in the actual climate record, v, defined in (2.2.7). For testing a single simulation, the statistic is defined as follows (SUN12, Appendix):

$$U_R = \frac{R\left(x_{\mathrm{f}}, v\right)}{\sqrt{\mathrm{Var}\left(R(x_{\mathrm{f}}, v)\right)}} \tag{A.1}$$

where

$$R\left(x_{\mathrm{f}}, v\right) = \frac{\sum_{t=1}^{n} \widetilde{w}_t (x_{\mathrm{f}t} - \mu_x)(v_t - \mu_v^{(\widetilde{w})})}{\sum_{t=1}^{n} \widetilde{w}_t^2 (v_t - \mu_v^{(\widetilde{w})})^2},$$

$$\mathrm{Var}\left(R(x_{\mathrm{f}}, v)\right) = \frac{\sigma_{\delta_{\mathrm{f}}}^2}{\sum_{t=1}^{n} \widetilde{w}_t^2 (v_t - \mu_v^{(\widetilde{w})})^2}$$

and $\mu_v^{(\widetilde{w})}$ estimated by the weighted average

$$\bar{v}^{(\widetilde{w})} = \frac{\sum_{t=1}^{n} \widetilde{w}_t v_t}{\sum_{t=1}^{n} \widetilde{w}_t}.$$

The correct expression for the weights $\widetilde{w}_t$ is given in Moberg et. al. (2015,p.427):

$$\widetilde{w}_t = \begin{cases} \sqrt{\frac{\sigma_\eta^2}{\sigma_\eta^2 + \sigma_\epsilon^2(t)}} & t \in \text{the reconstruction period} \\ \\ \sqrt{\frac{\sigma_\eta^2}{\sigma_\eta^2 + \sigma_\theta^2}} & t \in \text{the calibration period} \end{cases} \tag{A.2}$$

To arrive at a judgment of statistical significance of U_R one uses the fact that $U_R \overset{\text{approx.}}{\sim} \mathrm{N}(0,1)$ under the null hypothesis. Thus, if for example $U_R > 1.65$, then H_0 can be rejected at the one-sided 5% significance level. The higher U_R, the stronger the correlation between the model and the observations.

An example of using the *R* package sem

The model to be fitted is the 1st`o.i.FA(3,2)`-model with $\lambda_{11}, \lambda_{21}, \lambda_{22}, \lambda_{31}$, and λ_{32} as model parameters (see Table 1). Homoscedasticity is assumed.

```
mydata<-data.frame(x_f,z,x_total) # the data file consisting of
                                  # the values of the three
                                  # indicator variables

S<-cov(mydata)                    # the sample variance-
                                  # covariance matrix of
                                  # the indicators
library(sem)                      # to load the sem package

## Step 1:  model specification (here, in the path format)
model_FA32<- specifyModel()
F1 -> x_f          , lambda_11, NA  # a free param. with a start
                                    # value picked by sem itself
F2 -> x_f          ,         NA, 0
F1 -> z            , lambda_21, NA
F2 -> z            , lambda_22, NA
F1 -> x_total      , lambda_31, NA
F2 -> x_total      , lambda_32, NA
x_f <-> x_f        ,         NA, 0.0113
z <-> z            ,         NA, 0.0134
x_total<-> x_total ,         NA, 0.0134
F1 <-> F1, NA, 1
F2 <-> F2, NA, 1
F1 <-> F2, NA, 0
```

where `F1` stays for ξ'_{f}, while `F2` for $\xi'_{\mathsf{total}\perp\mathsf{f}}$. A single-headed arrow indicates a factor loading, whereas a double-headed arrow represents a variance or covariance.

```
## Step 2: estimation
options(digits=4)          ## set the number of digits in output

 #to specify heuristic measures of the model fit
options(fit.indices = c("GFI", "AGFI","SRMR","CFI"))

FA32<- sem(model_FA32, S, N=100,fit.indeces=TRUE)
                                  # where N is a number of obs.

# Completely standardized solution (see page 30, footnote nr.8)
coef(FA32, standardized=TRUE)
lambda_11 lambda_21 lambda_22 lambda_31 lambda_32
   0.7069    0.6739    0.5475    0.7272    0.5050
 # the solution is admissible
```

continued on following page

```
summary(FA32)
 Model Chisquare = 0.1913          # the G statistic  (2.3.20)
 Df = 1                            # with 1 degree of freedom
 Pr(>Chisq) = 0.6619               # and the associated p-value
 Goodness-of-fit index = 0.9987          # GFI  see (2.5.5)
 Adjusted goodness-of-fit index = 0.9923  # AGFI     (2.5.6)
SRMR = 0.005211                          # SRMR     (2.5.7)
 Bentler CFI = 1                         # CFI      (2.5.8)

 Parameter Estimates
           Estimate Std Error z value Pr(>|z|)
lambda_11 0.1063   0.01511   7.033   2.020e-12 x_f <--- F1
lambda_21 0.1572   0.02913   5.399   6.707e-08  z  <--- F1
    # reject H_0: lambda_21=0 at all conventional sign. levels
lambda_22 0.1278   0.03025   4.223   2.411e-05  z  <--- F2
lambda_31 0.1811   0.03029   5.980   2.237e-09 x_total <--- F1
lambda_32 0.1258   0.03438   3.659   2.534e-04 x_total <--- F2

Iterations = 12    ## the solution converged in 12 iterations

##    to calculate the ratio and
##    the associated 95% confidence set according to (2.3.25)
Lambda11<-coef(FA32)[[1]]
Lambda21<-coef(FA32)[[2]]
Var<-vcov(FA32)           # the matrix of the estimated variance
                          # and covariances among the estimates
var_Lambda11<-Var[1,1]
var_Lambda21<-Var[2,2]
cov_Lambda11_Lambda21<-Var[1,2]
a<-Lambda21^2-3.8415*var_Lambda21
a
  0.02166 # >0, which amount to rejecting H_0: lambda_21=0 at 5%
    sign. level
b<-Lambda11*Lambda21-3.8415*cov_Lambda11_Lambda21
c<-Lambda11^2-3.8415*var_Lambda11
r_1<- -sqrt((b/a)^2-(c/a))+(b/a)
r_2<- sqrt((b/a)^2-(c/a))+(b/a)
ratio<-Lambda11/Lambda21
result<-cbind(ratio,r_1,r_2)
result
         ratio     r_1    r_2
        0.6757  0.4308  1.1259
```

Since $a > 0$, the Fieller confidence region for the amplitude of a forcing effect in a climate model, represented by the ratio $\lambda_{11}/\lambda_{21}$, is a bounded confidence interval, i.e. $\Big[0.4308, 1.1259\Big]$.