DANISH METEOROLOGICAL INSTITUTE

——— TECHNICAL REPORT ———

03-37

Enhancing use of statistics in climate research: Transferring analysis methods that face up to commonly neglected data problems

An EU FP6 STREP outline proposal

Peter Thejll, Torben Schmith,
& FACEUP consortium

COPENHAGEN 2003

ISSN: 0906-897X (printed version)
1399-1388 (online version)

Specific Targeted Research Project

Enhancing use of statistics in climate research:
Transferring analysis methods that face up to commonly neglected data problems

FACEUP

Date of preparation : *15 October 2003*

Type of instrument : *STREP (Specific Targeted Research project)*

Submission stage : *OUTLINE proposal*

Activity code addressed : *NEST-2003-1 ADVENTURE*

Duration of the project (in months): *36*

Proposal abstract

Many statistical analysis methods currently used in climate research rely on the simplest of statistical assumptions. Advanced statistical methods take into account the failure of standard assumptions and their effects on results, and offer new techniques to overcome the problems. However, these up-to-date methods have not been widely embraced in climate research. Development of advanced statistical methods in the fields of econometrics and geostatistics has now surpassed the methods still used in climate research. This project aims to transfer knowledge of some of these advanced methods into climate science, because of expected gains in the reliability of statistical work. High impact on issues related to climate change policy is a possible outcome of the project. Benefits for the quality of climate research and thus society, in such areas as climate reconstructions, climate change attribution and downscaling, should result. Transferable statistical methods will be identified, along with climate problems ripe for investigation. A common model-based dataset has been chosen for analysis as the focus of the project. A toolbox of statistical methods will be distributed from an interactive project home page. The proposal does not fall within any FP6 thematic priority. In particular, it is not an ordinary climate research proposal, but is a novel trans-disciplinary approach intended to transfer knowledge, so that future climate research may benefit.

Introduction: At present, statistical methods in climate science do not, in the main, use methods from such fields as econometrics, geostatistics and time series analysis methods from theoretical physics, despite the problems under study having common characteristics. For example, the regression method based on 'least squares' is one of the most widely used techniques for estimating parameters in linear models in climate research. This method rests on tacit assumptions that are rarely stated or explicitly tested during use. Since the 1950's important improvements to regression methods have been developed and presented in fields outside climate science, and widely used in these fields. Unfortunately, they have not spread into climate research and become standard methods – they have only been applied sporadically. Appreciable improvements in estimating regression and other linear models would be available for climate research by the adoption of these new methods, and the aim of this project is to effectuate that transfer.

Main strategy in FACEUP: The project will **evaluate** several advanced statistical methods that have an impact on, for instance, ordinary least squares (OLS) regression methods – because OLS is one of the most widely used statistical methods in climate research, and any improvements that can be found and introduced will have a huge impact due to ubiquity of the method. At the heart of many statistical techniques lie the same statistical assumptions, and the proposed work can thus have an impact on several methods in use today.

The project will **identify** advanced methods that lend themselves to discipline-transfer, recognising that many advanced methods exist that are *not* easy to transfer. For instance, the 'Cochrane-Orcutt' method for doing regressions in the presence of serially correlated residuals is a method where the practitioner can easily see *how* the method works and understand *why* it can be used. Methods identified in the project will be turned into 'cook-book methods', as has been done with OLS for a long time.

Having identified advanced methods that are simple to appreciate and transfer between disciplines, the project will mount a major effort to **disseminate** these methods and the understanding of their benefits, to climate science. Changing people's attitudes is the core problem of the dissemination, and must be attacked with powerful means. In this work, showing what the method is, and highlighting the improvement it brings, is central. Repeatedly showing good applications, demonstrating them at scientific meetings, and discussing them in widely read papers and journals is one aspect of the approach. Organizing meetings on this specific subject is another. Additionally, presenting tools and papers on an **accessible Internet page** is important – the success of the Singular Spectrum Analysis and Multi-Taper Methods can in part be traced to the ready availability of an online 'Toolkit'. The spread of Wavelet analysis techniques can likewise be traced in part to the existence of an Internet site where people can download software and publications, but also upload datasets and receive wavelet analysis results back. This project will offer an interactive Internet site with downloadable software programmed in relevant platforms, and offer an upload-functionality for datasets that, for instance, are to be regressed and where both OLS and advanced–method results are returned.

Proposal novelty and ambition level: The suggested research is novel and ambitious because transfer of these practical, tested statistical methods into climate science, while attempted, has not 'caught on' on a wide scale – and the project seeks to perform this transfer *widely*. The project is complementary to existing efforts to transfer new methods between the field of *mathematical* statistics and climate science.

The Geophysical Statistics Project (GSP) at NCAR seeks to apply *standard* – although advanced - statistical methods to climate research.

The project is ambitious because a number of tacit assumptions commonly made in the analysis of climate data are being challenged. It is commonly assumed that processes in climate have short memory, and that standard regression assumptions apply (i.e., that residuals are 'white', or independent). This project considers methods that do not assume this and which offer readily available remedies. Introducing these methods to climate science may bring into question some commonly accepted results of, e.g. linear regression methods - such as climate reconstructions based on sparse proxy data; large-scale mean climatic series from the instrumental period; statistical downscaling; and climate change detection/attribution work.

Individual efforts to transfer econometric, or other advanced statistical methods, to climate science have been seen, but these remain isolated. An example of a successful idea that spread quickly is the use of principal component analysis (PCA) methods. Here, hundreds, if not thousands, of papers have been written in climate science using these methods, which have now become standard in the community. What this project wants to do is to transfer *other* easy-to-apply methods that are equally important, and spread the word to the climate research community. Incidentally, the PCA methods commonly rely on the very type of standard assumptions that the project challenges and will offer better methods for, so the potential impact of the basic project idea can be seen.

The scientific approach should be convincing because a group of leading experts from several communities intent on the knowledge-transfer has been formed. On the one hand are experts with profound experience of applying their methods to a variety of cases, such as: financial market dynamics, petroleum and mining exploration and biomedical signal analysis based on theoretical physics insight. On the other hand are climate scientists experienced in the statistical analysis of climate data. The project group consists of 13 people, four of which are women. The different concepts and methods will be tested in a common framework of extensive model datasets generated with climate models on two levels of complexity. The use of such a common dataset, and a working structure whereby statistical data analysis is performed - not in isolation - but in close collaboration and conference with climate experts, ensures that potential problems, such as nomenclature differences, are overcome. An extensive commitment to dissemination efforts in this project further underscores the clout of the proposal.

The proposal is plausible and the effort justified because the concepts and methods to be transferred have all been shown to work successfully within their fields of origin. The types of data encountered in the various fields are also quite similar. By identifying a number of 'test-bed' problems to be analysed with a selection of advanced methods it will be demonstrated that valuable knowledge transfer can occur. Intensive work on a joint dataset inside agreed-upon problem limits will give a common understanding of what sort of problems and opportunities lie in these data and methods. A common dataset based on long (1000 years long, highly resolved (twice daily), and 10,000 year long (monthly resolved)) dynamic climate model runs has been chosen, and the method-evaluation and -testing will be kept entirely within this data-world. Such data are of course only an approximation to reality but they are internally consistent and their complexity and multi-dimensionality ensures a thorough test of methods, so that transfer to 'real problems' can later be done based on this experience.

Project outputs include:

- Identification of possible caveats/pitfalls due to previous practise in determinations of climate reconstructions, climate change detection and attribution, and downscaling.
- Identification of methods to overcome these problems.
- Suggestions for changing statistical practise, the feasibility of these suggestions being demonstrated in the analysis of the model data sets.
- A dedicated Internet site providing software packages, an interactive data-analysis facility, and access to publications
- Extensive dissemination by arranging special sessions at climate meetings, writing scientific papers, monographs and arranging for 'special issues' of relevant journals, and by using research networks, such as PAGES, CLIVAR and EUROCLIVAR.

These objectives are realistic inside the 3-year duration of the proposed project because the participants are experienced within in their fields and therefore able to pinpoint relevant climate problems and appropriate methodologies. The team also has the necessary contacts to the scientific networks in order to disseminate the results to the broad community of climate research. Furthermore, the common pool of data is essentially available at the outset, thus avoiding a start-up delay, experienced within some other projects: existing model-output has only to be interpolated and formatted for use, and no further model runs are necessary.

State of the art: Few attempts to transfer this type of knowledge from applied statistics to climate science, on the scale considered, have been undertaken. Complementary approaches (the GSP) have mainly chosen to transfer new knowledge from the field of mathematical statistics, whereas the project seeks to transfer successfully tested applied methods already in use in other fields. The state of the art in 'ordinary' statistics is to assume that, for instance, residuals in regressions are un-correlated and that the Gauss-Markov theorem underlying ordinary least squares (OLS) therefore is satisfied. However, this is not always the case, and this may not be realized, but readily available methods for taking residual serial correlation into account are known in econometrics and geostatistics. Likewise, the role of 'long term memory' or 'integration/cointegration' in time series and its impact on such mainline use as OLS has not been widely considered in climate research. Econometrics has a panoply of standard methods available for use in cases where data commonly is not well described even by an AR(1) process – e.g. co-integration methods. This proposal is very timely – the 2003 Nobel Prize for economics has just been awarded to Robert F. Engle and Clive W. J. Granger who pioneered the econometric methods to be used in this project.

Assessing technical risk elements: Three main risk elements in the proposal have been identified:

1. *The new methods may not give superior results after all*

The data encountered in econometrics, geoscience, biomedical applications and climate science are not dissimilar, so improved results are expected when the more sophisticated methods are transferred to climate science. There are some *indications,* based on published work, that new methods may give improved results, but it needs to be checked via the suggested joint analysis of internally consistent model data, where, as one might say, 'the answers are known', and also checked against physical expectations about the climate system.

2. *It may not be possible to communicate across disciplines*

One aspect of this risk is *inside* the project group, but the funding requested will ensure the presence of adequate tools usually used to avoid this problem: **Frequent meetings and workshops** will be held in the project group, where people communicate and sort out nomenclature problems and, by collegial interactions, pave the way for common understanding of problems and methods. **Joint paper writing** is an even more intensive step in that direction, and the **composition of working teams** ensures communication across discipline borders, because statisticians and climate scientists will be involved in each of the working groups planned. Thus, no abstracted data analysis without a foundation in climate research reality should occur. During analysis of the model data the practical problem of not having a joint software platform to use for data analysis may be encountered, but this is simply remedied by ensuring access to such a platform, and a funding request for this is included.

Another aspect is the risk of failing to communicate the availability and superiority of the 'new advanced' methods to climate scientists *outside* the project group. However, the team will be actively communicating the results, not just in papers in high-profile journals, and 'special issues', but also by convening inter-disciplinary sessions at scientific meetings such as the annual EGS/EUG meetings, the AGU meetings, and smaller, more specific meetings. The participants will also spread news of what it is doing and accomplishing in summarizing vehicles such as the CLIVAR and PAGES publications which cater to various climate science communities that rely heavily on 'standard statistical' cook-book methods, such as OLS.

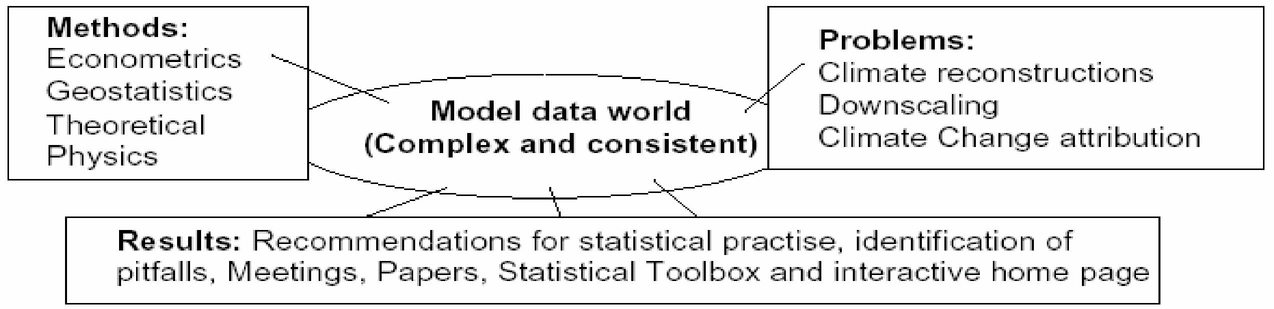
For use outside the project group, suitable tools programmed for the platforms commonly used in the climate research community will be constructed and distributed, so that there is no hindrance to adaptation and spreading the methods. Tools will be made readily available over the Internet. A funding request for this programming effort is included in the budget.

3. *Methods may not be transferable*

It may be that method limitations make the transfer to climate science impossible. For instance, the climate system may have 'regime change behaviour' or show other non-linear dynamics. If regime changes are frequent this problem is not inhibiting because methods can be adapted, but if climate regime changes occur only once or twice in an interval then the methods may fail – but this will be a joint problem for all methods, standard as well as advanced.

Risk/Benefit analysis: Thus, there are certainly failure risks, but the potential benefits are also high. The project could spread appreciably improved statistical methods to climate science. The benefits of this could impact upon many areas of research, such as more accurate estimates of global climate change and a quantitatively better understanding of its cause. These are important political and social issues for the EU.

Methodology: *The proposed project is a knowledge-transfer project* between certain fields of applied statistics and climate science. It encompasses tasks related to transferring knowledge, and tasks related to providing for the framework; that is, the model data which is to be the central focus of the project, and upon which the various methods will be tested and adapted. Several methods are to be introduced, as are some appropriate climate problems. The methods are those related to 'econometrics', 'geostatistics', and time series analysis based on theoretical physics insights, while the selected climate problems are those of reconstructions of historical climate; estimation of the role of external forcings; detection and attribution of climate change; and spatial problems (downscaling) that take correlations into account. The figure below schematically shows the project structure.


Methods:
Econometrics
Geostatistics
Theoretical Physics
Model data world (Complex and consistent)
Problems:
Climate reconstructions
Downscaling
Climate Change attribution
Results: Recommendations for statistical practise, identification of pitfalls, Meetings, Papers, Statistical Toolbox and interactive home page

Two groups in the project in possession of long coupled climate model runs will provide the model data for the analysis. These data have been analysed previously with conventional techniques to establish spatio-temporal patterns of variability and the response of the system to external forcings. This work involves extracting data of a predetermined type from a large body of model outputs already in existence.

The OLS method dominates as an estimation tool in climate science but improved alternatives exist. These alternatives include estimation of linear models with structured residuals, estimation of models relating non-stationary time series; and tests of time series for non-stationarity. Concepts from geostatistics are applicable in such questions as the formation of a global average temperature from unevenly distributed observing stations across the World. Commonly, series are averaged in boxes, with attention paid to the variations in decorrelation length across the World mainly for estimating uncertainties, rather than using an adapted method from the start. The project will study and compare the local characteristics of the model climate data. This can reveal which local area (e.g. land vs. sea) dominates the global scaling behaviour observed for the model data, and how local and global averaging affects the resulting global (or hemispheric) data. This will help in understanding the implications of the creation of reconstructed climate series, which are often based on local proxies. The response of the climate system to external forcings may be better characterized by methods that take, for example, the non-stationarity of the forcing into account, than by conventional linear regression techniques. In downscaling applications spatial correlation scales must be taken into account.

The work of ensuring knowledge is transferred between the communities involved will take place in the work itself, as explained above, but also handled by a specific dissemination task. All project participants will be assigned duties in the interest of dissemination, such as attending scientific meetings and presenting progress, establishing new collaborations, preparing materials for assessment reporting agencies, and to 'spreading the word' of the project goal and its potential benefits by writing peer-reviewed papers, and participating in 'special issue' preparations.

A professional management partner will handle those management tasks that can be handled without reference to the scientific work. The coordinator will handle scientific management tasks. An external review panel will be attached to the project.

RESOURCES

Total estimated project cost: 2771 k€

Total requested grant to the budget: 2282 k€

Partner No.	***Personnel Resources (man months)***	***Major Direct costs (keuro)***
RES-1	72	54 (*)
RES-2	36	
RES-3	36	
RES-4	18	
RES-5	18	
OTH-6	4	
RES-7	18	
RES-8	18	
RES-9	18	
RES-10	36	
RES-11	36	
RES-12	18	
RES-13	18	

(*) subcontractor for dissemination tasks

FACEUP consortium members:

Danish Meteorological Institute	Peter Thejll & Torben Schmith
Friedrich-Alexander Universität Erlangen-Nürnberg, Germany	Richard Reichel
Fachhochschule Ingolstadt, Germany	Jörg Clostermann
GKSS, Germany	Hans von Storch
University of Bern, Switzerland	Jürg Lutherbacher & Daniel Dietrich
Oxford University, UK	Myles Allen & Dáithí Stone
NIMH, Rumania	Aristita Busuioc
ACORE, Denmark	Lone Falsig
KNMI, The Netherlands	Nanne Weber
ARMINES, France	Hans Wackernagel
University of Rome, Italy	Giovanna Jona Lasinio
Giessen University, Germany	Armin Bunde & Jan Kantelhardt
University of Dortmund, Germany	Philipp Sibbersten