AnaEE
Infrastructure for Analysis and Experimentation on Ecosystems

Grant Agreement Number: 312690

SEVENTH FRAMEWORK PROGRAMME

CAPACITIES
RESEARCH INFRASTRUCTURES
COMBINATION OF CP & CSA

Milestone MS 12

Milestone title: Workshop on advanced computation tool for environmental data (scientific integration tools in ecosystem experiments).

Due date of milestone: 20 Actual completion date: 17

Start date of the project: November 1st, 2012 Duration: 42 months

Organisation name of lead contractor: Bioforsk

Contributors: Bioforsk, ITU, DTU, RRes

Revision N°: final

1. Achievement

The present milestone (MS 12) corresponds to the activity in task 3.3.2., from the DoW: “Proposing new advanced data analyses and computation tools for environmental data. One workshop will be organized on new computational tools.”

1.1 Introduction

The analysis of large amounts of data, often called “big data”, is promising to revolutionize many fields of science. Any research infrastructure that will be collecting a large amount of data in order to improve our power of analysis and our ability to address the Grand Challenges needs a strategy for extracting correct information and drawing accurate conclusions from the wealth of data being collected within the infrastructure. A key function of distributed research infrastructures is to allow us to conduct large-scale analysis across space, time and data complexity. For example, in earth and ecosystem science, the ICOS infrastructure aims at understanding the present state and to predict future behaviour of the global carbon cycle and greenhouse gases emissions based on extensive networks of standardized monitoring stations. As compared to ICOS, the ANAEE infrastructure aims at answering a greater diversity of issues linked to the impact of climate and other global changes on the services that ecosystems provide to society. Evaluation of adaptation strategies are at the core of ANAEE, and are encapsulated in the “experimentation” component of ANAEE: “Analysis and Experimentation on Ecosystems”. The flexible experimental framework of ANAEE is one of its strengths in answering the Grand Challenges, but it also calls for a strategy to extract relevant answers to these Grand Challenges based on multiple experiments conducted at local, regional , national or EU level.

Statistical modelling, including data fusion, provides key tools that allow us to conduct large-scale analysis across datasets. Any question asking at the larger scale if there is an effect and what are the drivers of this effect is a question that needs to be answered with statistical modelling. As an example, let us take the question of probing the link between soil biology and yields in Europe. First, we would need to conduct multivariate statistical analyses to explore the links between soil biology and yields in Europe. If we were to find such a relationship, we would further use multivariate analysis to see if this relationship is linked to climate and to soil management. Keeping the same hypothetical example, we might be able to isolate a key indicator of soil biology that is giving us the best relationship with yield. However, this indicator might be a time-consuming and expensive measurement to make. We would then be looking for parametrizing cheaper measurements that would act as a proxy to obtain a cheaper and faster estimate of the desired indicator. Again, the determination and parametrization of proxy sensors on large datasets is a task for statistical modelling. This simple example shows that proper statistical modelling is central to successfully answering the questions of an infrastructure that is by definition aiming at finding solutions across time scales, spatial scales and methods.

Our ability to conduct large-scale statistical modelling depends to a large extent on data accessibility and structure, therefore on the type of databases that data will be extracted from. Because data organization and analyses are intimately linked, we organized a workshop on "computational databases" for ANAEE, as a joint activity between WP3 and WP4. Here, we called "computational

databases" the development of databases and analytical algorithms, largely based on multi-variate statistics, for large and complex data sets. The fundamental hypothesis of our workshop was that multivariate processing of terrestrial ecosystem data across Europe (and beyond) will help us answer research questions far beyond meta-analysis of multiple published studies. The objectives of this workshop were first: 1) to determine the needs of the ANAEE research community and of ANAEE projects for multi-variate analyses, and 2) to assess what is realistically achievable, so as to 3) to map the key actions that would need to be implemented in the AnaEE infrastructure to support cohesive analysis of large amounts of data.

1.2. The Workshop

The workshop was organized from March 18th to March 19th 2014 on the Ås Campus, close to Oslo in Norway. The workshop was organised through a BaseCamp application for ease of communication.

Participants to the workshop were:

NAME	**AFFILIATION**	**COUNTRY**	**KEYWORDS**
ANNELENE PENGERUD	Bioforsk	Norway	MIR models
ATTILA NEMES	Bioforsk	Norway	Meta-analyses, databases in soil physics
BORIS JANSEN	Univ. Amsterdam	Netherlands	Biomarkers, models
CHRIS RAWLINGS	Rothamsted Research	UK	statistical genomics, databases, ANAEE
CHRISTOPHE MONI	Bioforsk	Norway	Soil fractionation (Standardisation)
CLAUS BEIER	DTU / NIVA	Denmark / Norway	Experiment networks (EXPEER, CLIMMANY), ANAEE WP3 leader
DANIEL RASSE	Bioforsk	Norway	Models, soil molecular data, ANAEE
ERIC COISSAC	LECA	France	bioinformatics
GUIDO WIESENBERG	Univ. Zurich	Switzerland	soil biogeochemistry; database structure
HARALD MARTENS	NOFIMA	Norway	multivariate statistics
JANNIS BUTTLAR	Max Planck Institute for Biogeochemistry	Germany	Biogeochemical cycles; data assimilation; data mining
LAURIC CÉCILLON	IRSTEA	France	Molecular databases and mutivariate statistics
LUIS RODRIGUEZ-LADO	Univ. Santiago de Compostela	Spain	statistical and spatial modelling for soil
MIKLOS DOMBOS	Hungarian Academy of Sciences	Hungary	Hungarian Soil Databases
NÜZHET DALFES	Istanbul Technical University	Turkey	Climatic databases, AanEE WP 4
PIERRE BARRÉ	ENS	France	European Long-term bare fallow network
SYLVIE QUIDEAU	Univ. Alberta	Canada	soil biogeochemistry

The programme was as follows:

Tuesday March 18th

Arrival from Hotel: 8:20

8:30 am	Welcome and Introduction to ANAEE	Daniel
8:45	Data Access Issues – AnaEE Project WP 4	Chris Rawlings
9:00	Questions about ANAEE	

9:15 Session: standardized protocols and common key dataset

9:15	Towards a standardization of warming experiments:	Claus Beier
9:25	Coping with soil fractionation	Christophe Moni
9:35	Integration of climatic data	Nüzhet Dalfes
9:50	Open access databases in soil physics	Attila Nemes
10:05	The MOLTER database	Lauric Cecillon
10:20	Break	
10:40	Data explorations in biogeosciences	Jannis Buttlar
10:55	The terradegra Hungarian soil database	Miklos Dombos

11:10 Discussion: defining the set of analysis tool needed for terrestrial research. Where are the greatest benefits expected?

12:00 Lunsj

13:00 Discussion (continued)

14:00 Session 3 : soil spectral data

14:00	Methods to process spectral data	Harald Martens
14:15	Specific challenges for MIR for soil	Annelene Pengerud
14:30	Statistical tool for Mir and NMR data	Lauric Cecillon
14:45	Statistics in the Cloud	Luis Rodriguez Lado

15:00 Discussion: standardization of spectral data acquisition and treatment for multivariate processing.

Wednesday March 19th

8:30	Session 4: soil biological and molecular data	
8:30	Bioinformatics: barcoding for soil life	Eric Coissac
8:45	Biomarkers	Boris Jansen
9:00	Need for Database in biogeochemistry	Guido Wiesenberg
9:15	Carbon biogeochemistry and soil biology	Sylvie Quideau
9:30	Key soil mineralogy measurements	Pierre Barré
9:45	Discussion.	
10:15	Break	
10:30	Dicussion Biomarkers (continued)	
11:30	Lunsj	
12:30	Session 5. Conceptual design of multivariate multi-location experiment	

1.3. Key findings

The results of the workshop will be further elaborated in the D3.2 “Models and new computational tools” due at month 30. Here. A preliminary summary of the findings is presented here.

The following key elements were identified:

1. Fairly standardized multi-sites and multi-location experiments appear to be the prime candidates for benefiting from computational database tools. Large data fusion schemes do not appear to hold much promise or have much appeal for the terrestrial ecosystem community at this stage, notably due to the extreme variability in data collection schemes, measurement methods, ...
2. Databases should be implemented at the project level (vs. a central database, where people send their data). AnaEE should therefore:
    a. develop a consistent database format and structure for all ANAEE projects.
    b. provide guidance and training for implementation of the database (including metadata) and archiving possibilities to ensure that the database will live longer than the project in ANAEE.

   c. provide database reporting tools for single-projects and multiple-projects, in order to extract data sets and make composite databases for analytical purposes.
   d. Simplify the process of submitting quality assured and "published" data together with metadata from individual AnaEE site or project databases to a central repository.
3. Provide help with multivariate statistics for AnaEE projects. The workshop clearly indicated that few scientists in terrestrial ecosystem research have sufficient statistical background to run complexe multivariate analyses themselves. Often, the scientists might not even see the possibility and generate experiments with sub-optimal design. Therefore AnaEE should provide help at several levels:
   a. Training courses in multivariate analyses.
   b. Advice in experimental design.
   c. Training in analysis of patterns of variability, which are often poorly treated in terrestrial research.
   d. Freeware open source analysis tools (such as R scripts) (community open) fully compatible and integrated with the read interface of the AnaEE database system.
4. Provide expertise on high performance computing for large-scale modeling.
5. Provide climate scenarios and parameters in ANAEE database formats. It was recognized that climate is a main driver in most ANAEE studies. Consistency and high quality for climate data (including future) and climate parameters are important. For example, an important question is how to extract from long-term climate time series (measured or predicted) relevant and statistically sound climate-sensitivity parameters. It is therefore suggested that the ANAEE computational group should include climate data analyses and formatting as support to ANAEE projects.
6. Standardization of methods. Non-standard methods are a major obstacle to conducting large-scale statistical analyses on existing data sets. Standardization is a priority consideration within project (e.g. multi-site projects). This contrasts with standardization across platforms / projects where it is very difficult to do, and often not necessarily desirable. The standardization should include the development of transfer functions between methods, and reference sample collections.
7. Foster projects making available published data in AnaEE format. This implies that AnaEE should also support projects that are exploratory in nature, and not only tightly hypothesis driven.

2. Next steps

2.1 Towards a set of recommendations

The findings of the workshop will be further developed in the D 3.2 "Models and new computational tools" due at month 30. These findings will be illustrated with key examples from the workshop.

Further, we will work towards sets of recommendations and guidelines to be presented in D3.2. These recommendations will largely be based on the findings of the workshop with respect to the needs of the AnaEE community and the realistic expectations in terms of “big data” analysis. In addition, recommendations will take into account solutions that are being considered and/or implemented by other large-scale projects for terrestrial research. It is expected that these solutions will be principally relevant to the database component of the “computational database” approach, and therefore a shared activity with WP4.

The present MS is with respect to task 3.3.2. “Proposing new advanced data analyses and computation tools for environmental data”. In D 3.2., we will make the link between the results of this workshop (and further work on task 332, as described above) with task 3.3.1 on “Task 3.3.1. Defining model and data requirements”.