ExArch: Climate analytics on distributed exascale data archives

M.N. Juckes (Corresponding author: martin.juckes@stfc.ac.uk), V. Balaji,
B.N. Lawrence, M. Lautenschlager, S. Denvil, G. Aloisio, P. Kushner, D. Waliser

1 Background

The climate modelling community faces huge challenges if it is to meet demands for quantitative predictions of risks associated with anthropogenic climate change. The precise roadmap to meeting these demands, yet to be drawn, is certain to push computing resources to their limits, to the exascale and beyond. While the problems with creating codes to run at the exascale are significant and need urgent attention, the associated problems of handling staggering volumes of globally distributed data cannot be neglected. This proposal concentrates on the research necessary in order to address some of those data issues as climate predictions become exascale.

Climate modelling is progressing in 3 principal directions: **(1) completeness** of the representation of the complex components of the Earth system and all their interactions; **(2) accuracy** with which inherent and "knowledge-gap" uncertainty is described; and **(3) improving precision** of the spatial representation of the physical world.

The Coupled Model Intercomparison Project Phase 5 (CMIP5) project[1], organised by World Climate Research Programme (WCRP) in order to provide the foundations of the 5th Assessment Report of the Intergovernmental Panel on Climate Change (IPCC), attempts to combine progress in all these direction in the best possible way. Compared with the inputs to the 4th Assessment Report (AR4), precision is improved by roughly halving the typical mesh size in the models, completeness is significantly advanced by including the carbon-cycle, and the accuracy with which uncertainty can be assessed is improved through a wider range of experiments. CMIP5[1] will produce of the order of 10 petabytes of data, from 60 experiments, each with multiple realisations by up to 40 different models (or model configurations).The data volume will increase 300-fold relative to that used in AR4. This rapid increase is associated with the increasing model resolution and complexity, but the increasing detail of the data stored is also a factor.

Nevertheless, the spatial resolution of the CMIP5 models (typically 150km, higher for some short term projections) will be too coarse for many climate impacts applications: dynamical or statistical downscaling is required. The CORDEX[2] project is designed to meet this requirement, producing downscaled data at resolutions from (initially) 50km increasing to 10km in many regions as the project progresses.

The data archive also faces a requirement for greater **timeliness** in the delivery of impact analyses. Typically, global model experiments are followed by a period of analysis and a delay for publication of results. The delay is then repeated for regional models before the work of impact assessment can be started, with the consequence that, as in AR4, impact assessment is often based on the projections from a previous generation of climate models. As the emphasis shifts from long term mitigation to short term adaptation, such delays are no longer acceptable.

These problems (compleness, accuracy, precision and timeliness) will only be exacerbated by the move

[1]cmip-pcmdi.llnl.gov/cmip5/index.html; Taylor et al, 2009 (tiny.cc/cmip5Design); Meehl and Hibbard, 2007, IGBP Report 57, WCRP: Geneva, 35 pp.

[2]COordinated Regional climate Downscaling Experiment: tiny.cc/cordex101

towards exascale and call for a comprehensive global approach to how scientists interact with peta-to-exa scale archives. While initial solutions are being put in place for CMIP5[3], the work carried out thus far has shown that there are major research problems to be addressed before those solutions will scale further as data volumes and global distribution increase alongside expectations for timely access. Eventually scientific users and application software will want to quickly grasp the content of datasets; formulate queries that may be dispatched to multiple physical archives; and detect features and patterns in coordinated experiments run across multiple modeling centres.

The ExArch proposal is principally a framework (incorporating a strategy, prototype infrastructure and demonstration usage examples) for the scientific interpretation of multi-model ensembles at the peta- and exa-scale. Specifically, we plan to do this in the context of the imminent CMIP5 archive, which will be largest of its kind ever assembled in this domain. We will further extend the challenge by attaching the ExArch framework to the CORDEX experiment, which will push even beyond CMIP5 in resolution, albeit on the regional scale.

2 The ExArch Team

The ExArch team consists of experts intimately involved with every step of the development and scientific interpretation of multi-model ensembles, has expertise in all the relevant areas: performing globally coordinated modeling experiments; development of metadata frameworks to standardize model output across the community; the science of regional climate; the statistical challenges in detection and attribution of climate change at global and regional scale. The team is already involved with efforts building the current state of the art in the field, including the Earth System Curator, Metafor, and Earth System grid projects, as wel as being in the thick of performing the CMIP5 experiments and constructing the CMIP5 archive. ExArch will join these national archives into a coherent whole and explore the added value associated with a comprehensive distributed archive.

The institutes contributing to the team have overlapping expertise, but within this project they will have a primary focus on the following roles: STFC, web processing services; Princeton, design of a query syntax for web processing; IPSL, extending and exploiting the common information model; DKRZ, portable processing operators and quality control; Toronto, scientific diagnostics; UCLA, satellite data for model evaluation.

3 Project challenges

The resolution of the atmospheres in global and regional models can be expected to increase towards 10km and 2km, respectively, by the end of the decade, continuing the rapid expansion in data volumes described above (by between two and four orders of magnitude). This project will exploit the CMIP5 experience handling multi-petabyte archives so that it will be possible to build multi-exabyte archives by the end of the decade. The project will work with infrastructure providers (some of whom are partners) to deliver real solutions as research tasks are solved. The infrastructure will have to provide timely, efficient, scalable, resilient and transparent access to geographically distributed archives run on heterogeneous platforms. Interoperability of components will be essential to the provision of flexible and chainable server-side processing. Running prototype systems with large operational datasets with active user communities will test

[3] Current solutions for CMIP5 are built around the U.S. Earth System Grid project. We are already working closely with the ESG team, and will continue to so under the auspices of this project. The relationship with ESG is discussed in the management plan.

resilience, while adherence to emerging inter-disciplinary standards will promote scalability. The project will engage with CORDEX as a real user group with data volume problems similar in scale to CMIP5.

Consider a query, of the kind that might be posed to an archive of the exascale future, requesting the projected frequency of intense tropical cyclones in some region of the globe for input into an impacts model. The projections of cyclone activity will need to be sampled across a probability distribution of outcomes from multiple climate realizations from different models, stored at different locations. The execution of the query will involve several steps that, in an exascale environment, will need to execute automatically:

- evaluation of provenance and quality control metadata to determine which datasets to include;
- despatch of queries to data nodes, negotiating authentication and access control layers;
- collection of results from the data nodes, evaluation of return codes for fault detection;
- further calculations to combine collected results;
- archive results for re-use;
- delivery of processed results to the end-user, perhaps in deferred fashion if the associated computation needs to be scheduled on a "cloud".

The provision of quality control information poses both informatics and climate science challenges: techniques for assessing uncertainty are evolving rapidly as data volumes grow. The informatics challenge is to develop a structured vocabulary to describe characterisations of uncertainty.

4 Research plan

The research will be organised into 3 work packages: (1) management and strategy development, (2) technology and 93) climate science and data validation. The central subject of this project is finding solutions which can be deployed in a real infrastructure to support scientific analysis of large distributed datasets. Many of the informatics problems which need to be resolved are already well characterised (if not solved), but it is essential that the solutions developed in the informatics context are tested against real problems of interest to scientists today. WP3 will provide this proof of utility, so that at the end of the project the infrastructure will be available together with a suite of analysis codes which exploit it to produce derived products of current scientific interest. WP3 will also provide quantitative quality control information, which will allow users of an archive with millions of datasets to avoid using data which has been published for validation and is not suitable for detailed analysis. This information will feed back into the detailed metadata systems established in WP2.

The work of the project is broken up into a series of tasks, in three series: “M” for management, “E” for enabling and “R” for research. The enabling tasks will provide basic infrastructure to support the research, using techniques which are sufficiently well known to be deployed in a semi-operational context. Each research task will produce a research publication, while management and enabling tasks will result in internal reports.

WP1: Management and strategy development.

WP Objective: To develop a strategy which can support a global infrastructure to support the use of the data products of exascale climate computing.

WP Context: Climate science is intrinsically global, both in subject and significance, and a globally coherent strategy is essential if exascale resources are to be used efficiently.

T:1 [M] Project coordination

Detailed in management plan.

T:2 [M] Strategy development for an exascale climate archive

We will split the strategy discussion into two phases; the first phase will cover hardware (including energy usage of different storage options) and the "transport layer" (e.g. gridFTP) and be completed (with publication of a strategy white paper) by the end of year 2 of the project. The document will lay out expected capabilities in low, medium and high investment scenarios. The second phase will discuss the expected user requirements and the services which could be provided given the transport scenarios identified by phase one. For each phase, the project partners will prepare a position paper to provide focus for a subsequent workshop (see management plan for details of workshops).

T:3 [M] Review & Participation in Governance structures

The effectiveness of an exascale earth system data distribution system will depend on efficient management and implementation of a wide range of standards and conventions. Two key conventions playing a fundamental role in CMIP5 which we would expect to be important for ExArch are the NetCDF CF conventions, and the METAFOR Common Information Model. Both may need extension and modification, and governance structures will be key to doing so. ExArch will be able to contribute to an international impetus to maintain the effectiveness of the governance of these key activities. Similarly, the community software development needed requires significantly more effort than has hithertoo been expended. A recent meeting under the auspices of ESGF (op cit) has begun to address such activities, and more meetings are expected under the auspices of the Global Organisation for Earth System Science Portals (GO-ESSP). ExArch will also play a key role in software governance particularly with respect to coodination of both development and deployment.

T:4 [M] Interact with GCOS Essential Climate Variables

The exascale climate model archive will require a high degree of interoperability with relevant archives of Earth Observation data, particularly those defined as “Essential Cilmate Variables”. This task will work with ESA, NASA and JAXA (and others as necesarry) to ensure that there is good communication between those developing the relevant data management plans.

T:5 [M] Evaluate requirements for global access to exascale resources

Earth system science is clearly not unique in its global reach, but it is perhaps unique in the requirement for end users to understand the scientific products. There is strong evidence that the impacts of climate change will be most strongly felt in some of the world's poorest regions. A thorough understanding of these impacts will not be obtained if scientists in these regions, with, for example, their local knowledge of human and bio-geophysical components of the food production system, do not have access to the climate projections. This task will evaluate the requirements of users in areas of poor network connectivity and the extent to which the software solutions of WP2 address those requirements. (ExArch will exploit other funded activities, including that of PI Balaji's Siebel Energy Grand Challenge grant from Princeton University to enable access to existing archives for South African students and scientists.)

WP2: Informatics research

WP Objective: Handling distributed exascale data will require a combination of improvements in both hardware (networks, disk systems etc) and software. ExArch will focus on the software requirements, while

staying in close contact with hardware developments[4]. Specifically it will address

A. Provenance and quality meta-data supporting the formulation of complex scientific queries across the distributed archive;
B. Fault tolerant collection and distribution of data and meta-data (inclusive of both faults of production, such as incompleteness, and of distribution, such as network and disk failures).
C. Unambiguous query syntax to support re-use and chaining of results;
D. Robust server side resource management;
E. Efficient distribution of large data volumes from multiple sources to multiple delivery points;
F. Interfaces which ensure security, transparency and inter-operability with Earth observation archives.

all in the context of the eventual necessity of a scalable and resilient infrastructure.

WP Context: ExArch will interact and/or exploit lessons from projects such as: Metafor and Curator, which provide the groundwork for model and output documentation; the European Grid Initiative (and similar activities elsewhere); and specific computer science activities such as those developing provenance metadata markup languages and algorithms for distributed processing. The project partners are familiar with many relevant projects, but it will be important to readdress the informatics landscape early in the project as this is fast changing technology area (and that is one of the challenges to adoption).

T:6 [R] Informatics Landscape

What projects can be exploited? Where are the outstanding technology gaps (not yet addressed in this proposal)? What are the actual expected data flows and how do they match the available networks? How will the massively parallel data streams be managed?

WP2.1 Software management

T:7 [E] Standardisation of software management

Reliable and re-usable software will be essential in an exascale environment. We will work with the ESGF Technical Working Group[5] to progress the devlopment of a software (and software deployment) strategy which will be able to support an archive distributed across hundreds of data centres and modelling centres, with highly heterogeneous hardware resources and differing levels of expertise. We will investigate the use of virtual machines in a cloud environment as a means of providing portability of services.

WP2.2 Robust metadata for exascale archives

T:8 [E] Collect metadata for CORDEX models

The CORDEX archive will provide a valuable test-bed for the systems developed in this project. This task will populate the METAFOR Common Information Model (CIM) for the CORDEX models, making any necessary extensions to the METAFOR CIM as required (while METAFOR as a project will conclude in mid-2011 it will leave appropriate governance structures for the ongiong evolution of the CIM).

T:9 [R] Automated generation of configuration metadata for ESMs

The metadata collection methodology for CMIP5 Earth System Models (ESMs) relies on information being entered into an online questionnaire. This procedure is susceptible to human error. Neither the element of human error nor the staff cost scale acceptably in an exascale archive. Scalable, automated configuration capture procedures are required. While METAFOR will investigate methods for doing this, robust

[4]Through established contacts with Alcatel and others.

[5]The Earth System Grid Federation technical working group has been set up under the auspices of GO-ESSP (see http://go-essp.gfdl.noaa.gov) to provide a venue for technical discussions to exploit the US Earth System Grid software in the development of a global infrastructure.

methodologies and solutions are not expected. The code management options used by a range of modelling centres (including GFDL, Met Office Hadley Centre, Max Planck Institute, University of Tokyo, and IPSL) will be reviewed and appropriate capture algorithms designed. A functioning capture algorithm implemented for two IPSL models (IPSLCM[6] (global, for CMIP5) and LMDZOR[7] (regional, for CORDEX)) will be delivered in year 3 of the project together with the underlying library.

T:10 [R] Automated translation of experiment configuration files

The design of any given climate simulation is captured in the CIM activity package, but the codes used to realise the simulation take the information from a set of configuration files and scripts. As in task R1, we will review the existing procedures for managing configurations files and then move on to explore the options for conversion from configuration file to CIM document and back. This will allow both rapid generation of CIM documents from existing configuration files and generation of configuration files from CIM documents.

T:11 [R] Extending the ESM Information Model to Earth Observation data

Earth observation data is, in many characteristics, different from climate model data. Nevertheless, there is a large overlap in the kind of information that a scientific data user will need to know about the data. The evaluation and exploitation of high volume climate data will be facilitated by using the same archive tooling for both climate data and earth observation data. This task will evaluate the potential for convergence between the CIM developed for climate model data and metadata models for Earth Observation data.

WP2.3 Query management in a exascale archive

T:12 [R] Query scope: define range of queries to be supported

The priority for an exascale archive will be to provide support for data intensive operations where performing these operations at or near the archive machines leads to a significant efficiency benefit. Two examples include finding exteremes and differencing:

Finding Extremes: The table below lists the steps in a calculation to evaluate as a function of latitude and month, the maximum daily mean temperature in a spatial domain covering Europe and in one decade from a set of data which spans a century. The first two step, which are highly generic, provide a 200-fold reduction in data volume and will already be supported by systems being deployed for CMIP5. ExArch will develop systems for further processing steps to achieve additional volume reduction.

Processing step	Volume reductions	Cumulative volume reduction
Extract temporal subset	10	10
Extract spatial subset	20	200
Extract longitudinal maxima	40	8000
Extract monthly maxima	30	240000

Differencing: Differencing two datasets clearly requires that they be transported to a single location, but differences and mean-square differences over sub-domains can be evaluated from the means and mean-squares over those domains. A sub-domain mean-square differencing query will provide a quick-look comparison of two large datasets without transferring the data. In many instances it will be necessary to transfer a mask evaluated from one or both datasets, but the data volume of the mask will typically be 32 times smaller than that of the full fields.

T:13 [R] A well formed Data Reference Syntax (DRS)

[6] www.ipsl.fr/en/Organisation/Horizontal-Structures/IPSL-Global-Climate-Modeling-Group
[7] lmdz.lmd.jussieu.fr/documentation/guides/lmdzor

Develop a standard for data reference (such as the CMIP5 DRS) and file/variable aggregation (e.g. CSML, CDML, NcML, THREDDS) that is able to describe multi-TB datasets in a form that is accessible to catalogues and data services and which supports chaining of processing requests.

T:14 [R] Data access from grid resources

Grid computing systems provide users with flexible access to geographically distributed computing resources: in order to support the task-to-machine allocation process, the exascale archive will need to provide accurate information on the result delivery latency to the candidate grid resources.

T:15 [R] Scalable query syntax implementation

Queries will be encoded according to an XML schema with a well defined serialisation into URIs. We will investigate the use of existing schemas such as MathML and Provenance ML to describe the chaining of processing operations, but this too is fast changing area [8]. In the above example, for instance, we may wish to calculate the maxima of a difference rather than the maxima of a field. The 4 operations listed would then have to be prefixed by a differencing operation. Service chaining is an essential requirement. ExArch will need to build on developing standards such as OGC Web Processing Service (WPS) and BPEL (Business Process Execution Language for Web Services).

WP2.4 Near archive processing

The concept of "server side processing" becomes vague when applied to a globally distributed archive. We envision an environment where data mining and data reductions services are run on a variety of machines with ultra-high capacity connections to the archive. The results are then passed on to users who may be using their local machines, or may be using a remote resource with a high-capacity link to the archive. This sub-package addresses improved methodologies for exascale data mining.

T:16 [R] Managing a distributed exascale data cache

As subsetting and server side processing are deployed to increase the efficiency of the archive, efficient caching of these derived products will be a new challenge. Example activities: (1) **Subsets:** This task will analyse pathways for dealing with caching of subsets. Suppose that request for land surface data over Western Europe is cached, and a subsequent request asks for the same data over the United Kingdom. What sort of a caching algorithm would have the intelligence to recognise that the UK is in Western Europe? More general factoring problems will be considered, and cost-benefit analyses will be conducted to determine optimal cache size and complexity. (**2) Commutation**: An efficient caching service should recognise when two expressions (e.g. *a*+*b* or *b*+*a*) yield the same result.

T:17 [R] Processing service

While there is a body of work on web processing services in both the Grid world and the OGC world, and these can be to some extent harmonised[9], there are a number of unanswered questions as to what would work best at exascale, and how to build effective services for users. The processing services will be modelled in UML, and a model archive, to simulate usage and failure modes, will be developed.

Such processing services would require three functionally independent units: a node, gateway and portal: **(1) The node** would deal with "atomic" requests which have no natural intermediate steps (for instance, each of the four steps in the extreme processing example above). We anticipate such a node providing an OGC

[8] Eg: http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Vocabulary_Mappings, as of August 2010.

[9] See Woolf and Shaon. An approach to encapsulating encapsulation of Grid processing within an OGC Web Processing Service, in *Grid Technologies for Geospatial Applications. GIS.Science, 3, 2009.*

interface to maximise interoperability, and it should manage a job queue and publish a queue status; **(2) The gateway** would deal with the logic of building complex requests from "atomic" requests; and **(3) The portal** would provide a "human friendly" user interface (UI) – access to all the functionality and to guidance information. ExArch will construct such a processing service to investigate how best to build such a modular service to support distributed climate data processing. A request building UI will be developed using a wizard approach, and using a virtual desktop (supporting a highly restricted shell script). "Cloud" and "grid" options for exploitation of distributed hardware resources will be evaluated.

T:18 [R] Security

In terms of securing these services, ExArch needs to confront and solve the problem of services acting as authorised agents in the service chain (i.e. so the use of an expensive resource can be requested by another resource, which itself interacted with the original authorising human and so on) . This problem has been revisited in nearly every Grid project, and is now becoming an issue in most complex web-service environments. ExArch will have to exploit a best-of-breed approach, and exploit it in a prototype to see if infrasructure providers will allow their resources to be exploited in this (international) manner.

WP3: Climate science and scientific quality control

WP Objective: A key goal of ExArch is to examine the issues associated with exascale climate computing and prototype real solutions with real scientific benefit. ExArch will promote the development of services to support CORDEX and further support CMIP5 with reference to Earth Observations (EO) from the JIFRESE[10] EO archive and re-analysis datasets. These services will be evaluated by carrying out a set of real scientific studies, addressing quantities of interest from a distributed archive using robust and scalable algorithms:

G. Evaluation of distributions of projected climate variables (for example, temperature, precipitation, windiness);
H. Evaluation of statistics for features (such as tropical and extra-tropical cyclones, extreme precipitation and drought) and essential climate variables;
I. Flexible specification of models and experiments used for ensemble calculations.

Characterisations of uncertainty will be included in G and H.

WP Context: Many of the possible solutions to handling exascale data may be theoretically attractive but founder upon ease-of-use and/or inter-institutional difficulties. The only way these can be assessed is by "whole system testing" on actual science problems. The problems selected here for road testing are also ones which push the boundaries of what can be done in terms of data, and are therefore suitable as guidelines as to what can be achieved beyond the end-of-the project with exascale data access. While increasing model resolution at exascale will lead to volume difficulties, so too will increased ensemble size and addressing uncertainty, and ExArch will focus on these within the climate science workpackage.

One of the outputs of this WP will be a library of command line operators (including the CDO library) which can be used in a command line mode. A selection of these will be implemented in the processing services developed in WP2, prioritising according to the benefits obtained by server side execution of each command. The documentation of these operators will enhance the reproducibility and transparency of climate analysis methods for stakeholders inside and outside the climate science community.

[10]The UCLA Joint Institute for Regional Earth System Science and Engineering (jifresse.ucla.edu)

WP3.1 Quality Assurance

T:19 [R] Scientific Quality Assurance – Schema Design

Quality control will become increasingly important in an exascale computing context. Researchers will be dealing with millions of data files from multiple sources and will need to know whether the files satisfy a range of basic quality criterea. For the CMIP5 archive a new 3 level quality control process is being deployed: level 1 is an elementary file syntax check. After passing level 1 files will be open to a limited group of people for additional checking. Level 2 applies some more sophisticated checks, and level 3 will include an expert review of a range of parameters. For the exascale archive we will need a system with greater flexibility and extensibility than this simple 3 level scheme. This task will address the first step: defining an XML schema to describe the quality assurance status of a file or file-set.

T:20 [R] Scientific Quality Assurance – Scope and Implementation

The scientific quality assurance should evaluate aspects of the data which can be computed objectively and unambiguously and which will support data selection decisions made by researchers. This task will identify a set of operations and implement them, complying with the schema developed in task 15. A digital signature will be used to provide reliable identification of the quality assurance provider.

T:21 [R] Climate Data Operators (CDO) in an exascale archive

CDO is a collection of command line Operators to manipulate and analyse Climate and forecast model Data[11]. A range of formats are supported and over 400 operators are provided. The current library is designed to work in a scripting environment with local files. This task will explore the extensions required to support efficient usage in an exascale archive with distributed data and computational resources. Operators which require non-trivial computational resources should be able to provide resource estimates to support scheduling decisions. Plain text output will need to be complemented with an extensible and self-descriptive form which supports aggregation of results from large file collections. This task will focus on the CDO library developments needed to support the evaluation of the diagnostics used in WP3.2 in an exascale archive.

WP3.2 Climate Science Diagnostics

A range of exemplar diagnostics will be used to stress the ExArch inspired infrastructure which address some key science questions. Tests will cover both the functionality of delivering results and the robustness and usefulness of responses to mis-formed or excessively demanding requests. Where necessary, additional collaboration (independently funded) will be sought to broaden the involvement of the climate science community. It is clear that the scientific content of different categories of simulation (centennial and short term global projections and simulations and regional downscaling results from CORDEX) contain different types of information and may require different analysis approaches[12].

T:22 [R] Consistency of models and observations

Observations will also be used in other tasks looking at a range of climate processes: this task will look at basic measures of consistency in climate fields. Both primary observations and re-analysis datasets will be exploited. Comparisons of mean fields and frequency distributions will be evaluated. Quantitative estimates of model error characteristics and bias corrections for the input to assessment models will be made. This task will exploit the validation databases prepared by UCLA and collaborators with independent funding[13].

T:23 [R] Monsoon systems and intra-seasonal variability in the tropics

[11] code.zmaw.de/projects/cdo/

[12] Knutson et al., Nature Geosciences, 3, 157-163, 2010.

Variability in monsoon systems may have a large societal impact, but models cannot predict it reliably[14]. The ability of the models to represent intra-seasonal variability such as the Madden Julian Oscillation and ENSO (El Niño-Southern Oscillation) will be assessed. A range of indicies will be evaluated and used to test the ability of the infrastructure to support evaluation across multiple simulation types.

T:24 [R] Atmospheric dynamics: cyclones, eddy-fluxes, modes of extratropical variability.

A range of cyclone identification and tracking algorithms[15] will be evaluated. *A* synthesis of the impact of climate change on cyclone frequency, intensity, and track characteristics will be created. Particular attention will be paid to problems associated with low model resolution and statistical corrections. Statistical analyses of the evolution of classical Eulerian storm track properties (geopotential height variance, eddy kinetic energy), blocking events, tropopause features, and extreme wind events will be evaluated across the CMIP5 and CORDEX ensembles. Extra-tropical eddy statistic variability will be analyzed in the context of modes of atmospheric low frequency variability (ENSO tele-connections, PNA, NAO/NAM, SAM, etc.) in the regional and global ensembles.

T:25 [R] Climate projections in seasonal snow cover, and implications for water resources

ExArch functionality for regional stakeholders will be tested on regionally focussed assessments of how changes to seasonal snow pack will affect climate and water resources. (1) University of Toronto, with research partners from Environment Canada and Natural Resources Canada, will investigate decadal-scale variability and trends in seasonal snow cover and snow water equivalent in the Canadian Arctic and Subarctic with reference to recent work on Arctic snow processes[16]. This analysis will examine statistics such as peak snow water equivalent amounts, snow melt onset dates and snow albedo feedback factors in Canada's North. (2) The UCLA group will use ExArch to develop a prototype system for seasonal and climatological water resources assessment in the Western United States by combining SWE observation and assimilation, regional climate model evaluation, and climate change impact assessment on water resources via snowpack and precipitation changes. UCLA will use ExArch to develop a prototype system for obtaining the seasonal and climatological data needed by water resources managers in the Western United States by combining SWE observation and assimilation, regional climate model evaluation, and climate change impact assessment on water resources via snowpack and precipitation changes. The data will be provided to California Department of Water Resources for water resources assessments through existing collaboration.

T:26 [R] Moist thermodynamics

The use of global and regional models to address the long-term variability and evolution of the global moist atmospheric general circulation will be evaluated[17]. The representation of systems which are heavily influenced by moisture will be evaluated in T24, 25. This task will look at moist thermodynamics at a more fundamental level of energy and entropy cycles, with a view to testing the ability of ExArch to provide researchers with tools that facilitate basic research.

---

[13] Lean et al. (2010 – report to WCRP): wcrp.ipsl.jussieu.fr/Workshops/RegionalClimate/Documents/2_5_Lean.pps; Teixera et al. (2009): http://www.wmo.int/pages/prog/wcrp/documents/Teixeira_et_al_Observations_for_IPCC_Sep_2009.pdf

[14] Randall et al., 2007: IPCC Assessment, WG1, Chapter 8.

[15] e.g. Wernli and Schweirz, 2006, JAS, 63, 2486-2507; Hanson et al., 2004, Climate Dyn., 22, 757–769; Dacre and Gray, 2009, Monthly Weather Review 137:1, 99-115.

[16] Wang et al. 2008, RSE 112(10):3794-3805, Zhao and Fernandes, 2009, J. Geophys. Res., 114, doi:10.1029/2008JD011272.

[17] Pauluis et al. 2008, Science, 321, 1075-1078; Laliberté and Pauluis, 2010, Geophys. Res. Lett. in press.

Annex 1: ExArch: Task effort allocation [staff months]

	Task	Total	BADC	DKRZ	IPSL	Princeton	Uni. Toronto	UCLA	Uni. Toronto (S)	CMCC*	Lead
	1 ***Project coordination***	**9**	9								BADC
	2 ***Strategy development***	**6**	1	1	1	1	1	1		1	BADC
WP 1	3 ***Governance structures***	**1**	0.5			0.5				1	Princeton
	4 ***Interact with GCOS***	**3**	1				1	1			BADC
	5 ***Evaluate requirements for global reach***	**1**	0.5			0.5					IPSL
	6 ***Informatics landscape***	**5.5**	2		0.5	3					Princeton
	7 ***Standardisation of software management***	**10**	2	2	2	2	2				DKRZ
	8 ***Collect metadata for CORDEX models***	**9**	1		8						IPSL
	9 ***Automated generation of the configuration metadata***	**11**		1	8	2					IPSL
	10 ***Automated translation of experiment configuration files***	**10.5**	1		7.5	2					IPSL
	11 ***Extending the ESM Information Model to EO Data***	**6**			3			3			UCLA
WP2	12 ***Query scope: define range of queries to be supported***	**13**		3		6	4				Princeton
	13 ***A well formed Data Reference Syntax***	**5**	1	2		2					Princeton
	14 ***Data Access from Grid resources***	**3**	1	1				1		1	CMCC
	15 ***Query syntax implementation***	**12**	2			10					Princeton
	16 ***Managing a distributed exascale data cache***	**5**	2			3					BADC
	17 ***Processing service***	**23**	8	4	4	3		4			BADC
	18 ***Security***	**6**	4		2						BADC
WP3	19 ***Quality Assurance – Schema design***	**10**	1	8		1					DKRZ
	20 ***Quality Assurance – Scope and implementation***	**9**		8			1				DKRZ
	21 ***Climate Data Operators in an exascale archive***	**8**		6			2				DKRZ
	22 ***Consistency of models and observations***	**11**					6	5			UCLA
	23 ***Monsoon systems and intra-seasonal variability .....***	**17**	2				6		9		Uni. Toronto
	24 ***Atmospheric dynamics***	**27**	2				10		15		Uni. Toronto
	25 ***Snow processes***	**18**						3	15		Uni. Toronto
	26 ***Moist thermodynamics***	**7**	1				6				Uni. Toronto

Summary

Task	Total	BADC	DKRZ	IPSL	Princeton	Uni. Toronto	UCLA	Uni. Toronto (S)	CMCC*
Months allocated by workpackage	**20**	12	1	1	2	2	2	0	
	119	24	13	35	33	6	8	0	
	107	6	22	0	1	31	8	39	
Total months allocated	**246**	**42**	**36**	**36**	**36**	**39**	**18**	**39**	
Number of tasks led		*6*	*4*	*4*	*5*	*4*	*2*		*1*
Own effort in tasks led by self		25	24	23.5	21.5	22	8		
Total effort in tasks led		52	37	31.5	36.5	69	17		

*	For CMCC (unfunded partner) a nominal value of 1 indicates significant participation.
Uni. Toronto (S)	There are two Uni. Toronto columns, the first for a PDRA, the 2nd for a student

World Climate Research Programme
c/o World Meteorological Organization
7 bis, avenue de la Paix – Case postale N° 2300
CH-1211 Geneva 2 – Switzerland

Ghassem R. Asrar, Director WCRP
Tel.: +41 (0)22 730 82 46
Fax: +41 (0)22 730 80 36
email: GAsrar@wmo.int

Geneva, 25 August 2010

Dear Sir/Madam,

I am writing in support of the proposal entitled "ExArch: Climate analytics on distributed exascale data archives" by M.N. Juckes, V. Balaji, B.N. Lawrence, M. Lautenschlager, S. Denvil, G. Aloisio, P. Kushner, and D. Waliser. The proposed research and the resulting capabilities and services that this proposed project will provide to the climate research community and users of climate information will be invaluable to the European scientists, and a very wide network of scientists around the world.

This Project will uniquely provide easy, reliable and timely access by the European scientists to the climate models simulation results that are being organized by the World Climate Research Programme (WCRP) working group on coupled modeling with active participation of all major modeling centers around the world. This is a unique arrangement to ensure highest quality of standards and stewardship in design, development, production and analysis of the envisioned data sets. The IPCC and other environmental assessments depend on these results, hence they are made available to all researchers and users around the world without any restriction. The support for this entire activity is provided through the invaluable national and multi-national contributions because of its significant importance, and tremendous impact globally.

Your favourable consideration and support for this proposed project will undoubtedly reflect on continued leadership and support of Europe and European scientists in this regard. WCRP is truly proud and privileged to be a part of this international partnership and recommend strongly your positive consideration of this proposal. I will be prepared to serve in a scientific and technical advisory capacity to this project if it is funded.

I will be pleased to provide a more detailed statement of support if needed.

Yours faithfully,

Ghassem R. Asrar
Director