Global hydrology modelling and uncertainty: running multiple ensembles with a campus grid

BY SIMON N. GOSLING1,*, DAN BRETHERTON2, KEITH HAINES2 AND NIGEL W. ARNELL1

1Walker Institute for Climate System Research, Agriculture Building, University of Reading, Earley Gate, Reading RG6 6AR, UK
2Environmental Systems Science Centre (ESSC), Harry Pitt Building, University of Reading, Earley Gate, Reading RG6 6AL, UK

Uncertainties associated with the representation of various physical processes in global climate models (GCMs) mean that, when projections from GCMs are used in climate change impact studies, the uncertainty propagates through to the impact estimates. A complete treatment of this ‘climate model structural uncertainty’ is necessary so that decision-makers are presented with an uncertainty range around the impact estimates. This uncertainty is often underexplored owing to the human and computer processing time required to perform the numerous simulations. Here, we present a 189-member ensemble of global river runoff and water resource stress simulations that adequately address this uncertainty. Following several adaptations and modifications, the ensemble creation time has been reduced from 750 h on a typical single-processor personal computer to 9 h of high-throughput computing on the University of Reading Campus Grid. Here, we outline the changes that had to be made to the hydrological impacts model and to the Campus Grid, and present the main results. We show that, although there is considerable uncertainty in both the magnitude and the sign of regional runoff changes across different GCMs with climate change, there is much less uncertainty in runoff changes for regions that experience large runoff increases (e.g. the high northern latitudes and Central Asia) and large runoff decreases (e.g. the Mediterranean). Furthermore, there is consensus that the percentage of the global population at risk to water resource stress will increase with climate change.

Keywords: campus grid; high-throughput computing; global hydrology modelling; water resources; climate change uncertainty; climate change impacts

1. Introduction

e-Science methods have had a fundamental impact on the way in which climate science is undertaken and have changed the very science questions that can be posed (Kerr 2009). The aim of this paper is to demonstrate how the application of one particular e-Science method, high-throughput computing (HTC), has benefited the environmental science modelling community by allowing us to answer specific questions about the degree of uncertainty in estimates of climate

*Author for correspondence (s.n.gosling@reading.ac.uk).

One contribution of 15 to a Theme Issue ‘e-Science: past, present and future II’.
change impacts. HTC is often carried out on distributed computing resources, i.e. computers that are geographically separated from each other to some degree. For example, the climateprediction.net project used the distributed computing resources of the general public to run ensembles of a climate model to examine the implications of raising carbon dioxide levels in the atmosphere (Stainforth et al. 2005). Here, we use HTC in a campus grid environment in order to estimate uncertainty in the impacts of climate change. We present results from an ensemble of global river runoff simulations and water resources stress indicators under different climate change scenarios, which have been enabled through HTC. The Natural Environment Research Council (NERC) QUEST-GSI project (http://quest.bris.ac.uk/research/themes/GSI.html) aims at a systematic investigation of how uncertainty associated with climate modelling propagates through to the climate change impact estimates of river runoff and water resources stress. To adequately represent the full extent of uncertainty due to climate modelling, an ensemble of 189 hydrological simulations was necessary in order to span the range of climate projections from 21 climate models and nine different scenarios of global warming. The results will be useful to policy-makers in determining the adaptation and mitigation options for dealing with the impact of climate change on river runoff and water resources stress. This large number of simulations was not practicable given the way that the hydrological impacts model was designed to run on a single personal computer (PC). Instead, the 189 simulations have been enabled by the University of Reading Campus Grid; however, a number of modifications had to be made to the global hydrological model (GHM) and to the Campus Grid in order for the simulations to be run in HTC mode. Here, we present the details of these challenges, which are generic and would apply to any similar climate impacts modelling exercise, along with the results of the GHM simulations.

In §2, we provide some background on the nature of the uncertainties explored here. In §3, we describe the hydrological and water resources impacts models, followed by details of the experimental design in §4. Section 5 describes how the impacts models were adapted for running on the Reading Campus Grid, and §6 shows how the Campus Grid itself was adapted to accommodate the impact models. The main results are presented in §7, and we finish by describing how these impact models are now being applied in further studies that use the Campus Grid (§§8 and 9).

We aim to present this paper in order to appeal to both the campus grid computing community and the environmental modelling community, and to demonstrate how collaboration between the two communities can result in a unique and significant contribution to environmental science that improves our understanding of the role of uncertainty in climate change impacts assessment, and ultimately plays a role in the policy-making process. It is hoped that the details and discussion presented here will encourage further such collaborations.

2. Uncertainty in climate change impacts assessment

The Fourth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC AR4) acknowledges that future climate change impacts research should demonstrate and communicate the degree of uncertainty that is inherent
Global hydrology modelling uncertainty

in climate change impact estimates (Parry et al. 2007). Climate change impacts are typically estimated by applying climate projections from a global climate model (GCM) to an impacts model such as a GHM. Importantly, the GCMs used to create projections of future climate introduce a source of uncertainty into the impact estimates (Gosling et al. 2009). GCMs typically represent the atmosphere, ocean, land surface, cryosphere and biogeochemical processes, and solve the equations governing their evolution on a geographical grid covering the globe. Some processes are represented explicitly within GCMs, large-scale circulations, for instance, whereas others are represented by simplified parametrizations. The use of these parametrizations is sometimes due to processes taking place on scales smaller than the typical grid size of a GCM (a horizontal resolution of between 250 and 600 km) and sometimes due to the current limited understanding of these processes. Different climate modelling institutions will use different plausible representations of the climate system, which is why climate projections for a single greenhouse gas emission scenario will differ between modelling institutes. A method of accounting for this so-called ‘climate model structural uncertainty’ is to use a range of climate projections from ensembles of plausible GCMs, to produce an ensemble of impact projections for comparison (Kerr 2009). Such crucial uncertainties have not been fully considered until recently in climate change impacts assessments, with some studies choosing to explore the impacts associated with only one or two GCMs (Bergström et al. 2001; Kamga 2001). This is described as a dangerous practice by Hulme & Carter (1999), as it could be misleading, especially with regard to the policy-making process. Therefore, the ensemble approach is applied here by using climate projections from 21 of the 23 GCMs included in the World Climate Research Programme Third Coupled Model Intercomparison Project (WCRP CMIP3; Meehl et al. 2007).

3. Hydrological impact models

GHMs model the land surface hydrological dynamics of continental-scale river basins. Here, we apply one such GHM, Mac-PDM.09 (‘Mac’ for ‘macro-scale’ and ‘PDM’ for ‘probability distributed moisture model’). Mac-PDM.09 simulates runoff across the world at a spatial resolution of 0.5 $\times$ 0.5 deg. A detailed description and validation of the model are given by Gosling & Arnell (2010). In brief, Mac-PDM.09 calculates the water balance in each of 65 000 land surface 0.5 $\times$ 0.5 deg cells on a daily basis, treating each cell as an independent catchment. It is implicit in the model formulation that these cells are equivalent to medium-sized catchments, in other words between 100 and 5000 km$^2$. River runoff is generated from precipitation falling on the portion of the cell that is saturated and by drainage from water stored in the soil. The model parameters are not calibrated—model parameters describing soil and vegetation characteristics are taken from spatial land cover datasets (FAO 1995; de Fries et al. 1998). Mac-PDM.09 requires monthly input data on precipitation, temperature, cloud cover, wind speed, vapour pressure and number of wet days (daily precipitation greater than 0.1 mm) in a month.

Mean annual runoff simulated by Mac-PDM.09 can then be used to characterize available water resources using the water resources model described in Arnell (2004). It is necessary to define an indicator of usage pressure on water

Phil. Trans. R. Soc. A (2010)
resources for the model. Here, we use the amount of water resources available per person, expressed as cubic metres per capita per year. The water resources model assumes that watersheds with less than 1000 m\(^3\) per capita per year are water stressed. Therefore, populations that move into this stressed category are considered to experience an increase in water resources stress. However, some populations are already within the water-stressed category, because present-day resources are less than 1000 m\(^3\) per capita per year. Therefore, a more complicated measure combines the number of people who move into (out of) this stressed category with the number of people already in the stressed category who experience an increase (decrease) in water stress with climate change. The key element here is to define what characterizes a ‘significant’ change in runoff and hence water stress. The water resources model assumes that a significant change in runoff, and hence water stress, occurs when the percentage change in mean annual runoff from Mac-PDM.09 is more than the standard deviation of the 30-year mean annual runoff due to natural multi-decadal climatic variability. Hence, the water resources model calculates the millions of people at increased risk of suffering water resources stress during climate change as the sum of the populations that move into the stressed category (resources less than 1000 m\(^3\) per capita per year) and the number of people already in the stressed category who experience a significant increase in water stress. By summing runoff across grid cells, we calculate catchment-level indicators of water resources stress, which are then aggregated to the global level. The indicators are based on the assumptions of projections of population and economic development for the 2080s (i.e. 2070–2099) consistent with the IPCC Special Report on Emissions Scenarios (SRES) A1B, A2 and B2 scenarios (Nakićenović & Swart 2000).

4. Sampling climate model structural uncertainty

To adequately explore the effect of climate model structural uncertainty on global runoff and water stress, we investigated the impact associated with climate change scenarios from 21 of the 23 GCMs included in the WCRP CMIP3 (Meehl et al. 2007). The spatial resolution of a GCM is coarse compared with that of the hydrological processes simulated by Mac-PDM.09, which operates at a 0.5 × 0.5 deg resolution. For example, the UK is covered by only four land cells and two ocean cells within the UKMO HadCM3 GCM. As such, the GCM output is generally not considered of a sufficient resolution to be applied directly in hydrological impact studies, and so the data need to be ‘downscaled’ to a finer resolution. Here, the climate change scenarios applied to Mac-PDM.09 were not the original output from GCMs, but were instead created using ClimGen (Mitchell & Osborn 2005), a spatial climate scenario generator that uses the pattern-scaling approach (Mitchell 2003) to generate spatial climate change information for a given global-mean temperature change from present and a given GCM. ClimGen includes a statistical downscaling algorithm that calculates climate change scenarios for a fine 0.5 × 0.5 deg resolution, taking account of higher-resolution surface variability in doing so. ClimGen was developed at the Climatic Research Unit at the University of East Anglia (UEA), UK. The pattern-scaling approach
Global hydrology modelling uncertainty relies on the assumption that the pattern of climate change (encompassing the geographical, seasonal and multi-variable structure) simulated by GCMs is relatively constant (for a given GCM) under a range of rates and amounts of global warming, provided that the changes are expressed as change per unit kelvin of global-mean temperature change. These normalized patterns of climate change do, however, show considerable variation between different GCMs, and it is this variation that ClimGen is principally designed to explore. Therefore, the application of climate change scenarios from ClimGen is ideally suited to exploring the effects of climate model structural uncertainty on climate change impacts.

Climate change patterns for each of the 21 GCMs associated with prescribed increases in global-mean temperatures of 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 4.0, 5.0 and 6.0°C (relative to 1961–1990), i.e. nine patterns for each GCM, were applied to Mac-PDM.09. Mac-PDM.09 can only be run with input data from one GCM pattern for one increase in global-mean temperature at a time, which takes around 3 h to complete on a typical single-processor PC. The ClimGen output files that Mac-PDM.09 uses for each simulation are around 0.8 GB and normally take approximately 1 h to create for a single GCM pattern and increase in global-mean temperature. They are normally downloaded from UEA and need to be reformatted to be compatible with Mac-PDM.09. Therefore, the combined total run-time for one GCM pattern and global-mean temperature increase is around 4 h (excluding the time taken to download the data from UEA). Given the potentially large number of simulations necessary, a prioritized list of simulations was formulated (table 1). The highest priority set, for the HadCM3 GCM, equates to 36 h of simulation run-time (9 prescribed temperatures × 4 h per simulation × 1 GCM). The entire set of simulations necessary to adequately sample climate modelling uncertainty equates to around 32 days of continuous simulation run-time (9 prescribed temperatures × 4 h per simulation × 21 GCMs). Although it was practicable to run up to the third priority set of simulations on a single PC, which would consider the uncertainty arising from seven GCMs (e.g. Prudhomme et al. 2003), the complete set of 189 simulations was not feasible.

It is the combined limitation of (i) the coarse resolution of GCM data, (ii) the long simulation times required, and (iii) the large amounts of data storage required, that no previous studies have explored the global impact of climate change on river runoff and water resources with an ensemble of GCMs as large as 21, which is required in order to adequately address the issue of climate model structural uncertainty. For instance, Vörösmarty et al. (2000) considered two GCMs in their assessment of the impact of climate change on global water resources. Although several studies have applied more GCM projections at the local scale for specific local river catchments—e.g. UK river catchments (seven GCMs; Prudhomme et al. 2003), the southern Alps (three GCMs; Barontini et al. 2009) and Western Australia (Charles et al. 2007)—none has so far applied this over the entire globe. The application of a geographically gridded GHM here means that we were able to consider not just a few, but over 65 000 catchments across the entire land surface of the globe. This has been enabled by adapting Mac-PDM.09 and ClimGen for large ensembles and for compatibility with the University of Reading Campus Grid.
Table 1. Mac-PDM.09 simulations necessary to explore the effects of climate model structural uncertainty on global hydrology. Each cell denotes a single Mac-PDM.09 simulation, which takes 4 h to complete. Numbers within each cell denote the original priority given to each simulation (1, highest).

<table>
<thead>
<tr>
<th>GCM pattern</th>
<th>0.5</th>
<th>1.0</th>
<th>1.5</th>
<th>2.0</th>
<th>2.5</th>
<th>3.0</th>
<th>4.0</th>
<th>5.0</th>
<th>6.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>UKMO HadCM3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CCCMA CGCM3.1 (T47)</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>IPSL CM4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>MPI ECHAM5</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>NCAR CCSM3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>UKMO HadGEM1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>CSIRO MK3.0</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>CCSR MIROC3.2 (HI)</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>CCSR MIROC3.2 (MED)</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>CNRM CM3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>GFDL CM2.1</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>GISS EH</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>GISS ER</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>INM CM3.0</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>MRI CGCM2.3.2</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>GFDL CM2.0</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>NCAR PCM1</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>BCCR BCM2.0</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>CCCMA CGCM3.1 (T63)</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>GISS AOM</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>CSIRO MK3.5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

5. Modifying Mac-PDM.09 for large ensembles

To run the full set of 189 Mac-PDM.09 simulations presented in table 1 would require downloading around 150 GB of ClimGen output data (0.8 GB x 189 simulations) from UEA, but this would not have been practicable. For instance, if modifications or updates were made to the ClimGen code that affected the ClimGen output, the entire 150 GB would need to be downloaded again. Furthermore, the Reading Campus Grid does not presently include the facility to enable a running model to download or read the data it requires directly from a remote data repository at another institution. Therefore, the ClimGen code was first obtained from UEA and integrated as a module within Mac-PDM.09, resulting in the ‘ClimGen/Mac-PDM.09’ integrated model. This meant that the ClimGen output files required by each Mac-PDM.09 simulation presented in table 1 could be produced by Mac-PDM.09 on demand at Reading rather than having to be downloaded from UEA. The input/output (I/O) routines of the integrated model were tailored for running the model on the Campus Grid, such that a hydrological simulation can be performed for any scenario of global-mean temperature change and GCM. Modifications mean that, instead of taking 756 h to run the simulations presented in table 1 on a single PC, the Campus Grid
Global hydrology modelling uncertainty

has enabled them all to be run in under 24 h. Several important modifications needed to be made to the Campus Grid first, and these are detailed in the following section.

6. Campus Grid modified for large I/O

The Reading Campus Grid is a Condor pool, a research computing grid containing several hundred library and laboratory computers that are made available for running scientific applications when they would otherwise be idle (i.e. unused) by the Condor distributed resource management system (http://www.cs.wisc.edu/condor/). Condor facilitates HTC, where multiple instances of the same application run simultaneously on separate processors, allowing a large number of computing tasks to be completed in a short space of time. This list of simulations presented in table 1 was therefore ideal for running on the Campus Grid, given that 189 4 h simulations were necessary for the required analysis. A set of simulations like this can be described as a parameter sweep. One Campus Grid computer was required for each of the 189 simulations. Therefore, the maximum number of computers that were used simultaneously for ClimGen/Mac-PDM.09 simulations at any one time was 189. Some scientists run this type of parameter sweep on high-performance computing resources such as national supercomputers and departmental and institutional clusters, but these were not immediately available to us, and, in any case, such an expensive parallel capability resource should not be necessary for single-processor applications such as described here.

The main challenges associated with running the ClimGen/Mac-PDM.09 integrated model on the Reading Campus Grid via Condor were the large data storage and transfer volumes involved. A single simulation presented in table 1 requires storage for 20 GB of ClimGen input files, around 0.8 GB of ClimGen output files and 55 MB of Mac-PDM.09 output. Therefore, the total amount of storage required for the 189 simulations was around 180 GB (20 GB for the ClimGen input files, 150 GB for the ClimGen climate output files and around 10 GB for the Mac-PDM.09 runoff output). None of these data were stored on the local disks of the computers running the models (usually called the compute nodes); instead, bytes of data were transferred on demand to and from the Grid’s main server (called the head node) as required by the models, via Condor’s remote system call mechanism.

The Grid’s head node has very little storage capacity of its own; therefore, a temporary ‘scratch’ storage partition on another server is mounted on the head node via NFS (Network File System) for use by Grid applications. However, the scratch partition is limited in size and shared by all users of the Grid, which means that the availability of 180 GB of free space required for 189 ClimGen/Mac-PDM.09 model runs could not be guaranteed. Therefore, it was necessary to find a storage solution that enabled running models to directly access data on larger, permanent storage resources belonging to the ClimGen/Mac-PDM.09 modelling group. Writing shell scripts to control file transfer between the permanent storage server and the compute nodes would be difficult and error-prone because of the complicated directory structure normally used to store the data. Instead, the permanent storage partition was mounted on the head node using the SSH Filesystem (SSHFS—http://fuse.sourceforge.net/sshfs.html), a filesystem client.
based on the SSH (Secure Shell) file transfer protocol. Compute nodes were able to access the permanent storage partition via Condor’s remote system call mechanism (using Condor’s Standard Universe execution environment), in the same way as they would normally access the NFS-mounted scratch partition. This allowed the carefully arranged model input/output directory structure to be retained in the Grid-enabled ClimGen/Mac-PDM.09 model setup, greatly reducing the effort required to adapt the model for use on the Grid. This data management arrangement is illustrated in figure 1.

SSHFS was chosen as the remote data access mechanism for the following reasons: (i) it is easy to use, (ii) it requires relatively little system administrator assistance to install and the server side software it uses (usually OpenSSH) is installed as standard on most Linux and UNIX servers, and (iii) in many situations, it provides ready-made, direct access to remote data through institutional firewalls, although firewalls were not an issue in this particular project. The ease of use of SSHFS is illustrated by the fact that a directory on a remote file server can be mounted locally by an unprivileged user with a single Linux command. We were looking for a solution that would be easy to implement by scientists (without specialist assistance) for other models and for other campus grids, as well as one that would help with the ClimGen/Mac-PDM.09 model runs. Network file systems requiring system administrative intervention to mount a remote dataset (such as NFS or CIFS (Common Internet File System)) or to reconfigure the remote storage and data export mechanism (such as AFS (Andrew File System)) are not appropriate in the case of a single scientist needing to perform a set of model runs on a particular dataset on a one-off basis. However, a disadvantage of SSHFS is the inbuilt data encryption, which is really unnecessary for this application.

SSHFS was able to cope with a maximum of 60 ClimGen/Mac-PDM.09 model instances running at the same time. Increasing this number led to unacceptably long model run-times accompanied by excessive computational load on the Grid server, which is likely to have been a combination of Condor load, due to periodic

Figure 1. Data management during ClimGen/Mac-PDM.09 model runs on the Reading Campus Grid. The modelling data store is mounted on the Campus Grid server via SSHFS, and the running models access these data via Condor’s remote system call mechanism. The Campus Grid’s own, limited file storage capability is not used.
checkpointing and data transfer between the Grid server and the compute nodes, and SSH-related load as a result of the encrypted data transfer between the Grid server and the data server. There was not enough time to conduct tests that were sensitive enough to identify the exact cause of the bottleneck, so instead a limit of 60 simultaneous runs was imposed by implementing a Condor group quota for the Campus Grid users concerned, which allowed all 189 jobs to be submitted at
the same time, with only 60 being allowed to run at any one time. It was possible to compare the SSHFS performance with NFS by using the Campus Grid scratch disk, which was described earlier. On one occasion, we were fortunate enough to have access to enough storage space on the Grid’s scratch partition for all 189 ClimGen/Mac-PDM.09 model runs, which enabled all 189 model runs to be carried out simultaneously while accessing the data via NFS instead of SSHFS. In the NFS test, all 189 runs were completed in less than 9 h. Thus, provided that the 60 job limit was imposed when using SSHFS, there was very little time overhead compared with NFS-managed jobs.

It was difficult to predict how long a full set of 189 runs would take to complete, even when the average run-time for an individual model instance was accurately known, because the use of the university’s library and laboratory computers varies from day to day and through the day, and therefore exhaustive tests to determine SSHFS and NFS performance were not conducted. Nevertheless, the experiments that were conducted allowed us to develop a practical data management solution for ClimGen/Mac-PDM.09 involving SSHFS and the Condor group quota mechanism, which was capable of completing 189 model runs in less than 24 h.

7. Results

Figure 2 (left panels) displays the 21-member ensemble mean (i.e. the mean across all 21 GCMs) percentage change in mean annual runoff for different degrees of prescribed global-mean temperature rise from present. The simulated changes in mean annual runoff are spatially heterogeneous. For instance, for any given global-mean temperature increase, the largest increases in runoff are observed in eastern Africa, north-western India, Canada and Alaska. Regions that experience the greatest decreases in runoff relative to present include southern Europe, South Africa, central South America, mid-western USA and Turkey. As the level of global-mean warming increases from 0.5°C to 6°C, the magnitudes of these changes increase. Importantly, the geographical extent of some of the regions that experience change increases; for instance, at 0.5°C, northern and western India display increases in runoff, with eastern India showing no change, but at 6°C, all of India presents an increase in runoff. Similar patterns are observed with the drying for eastern and southern Europe and the increases in runoff for Canada.

Quantification of the uncertainty in the ensemble mean of mean annual runoff changes presented in the left column of figure 2 can be taken from the right panels of figure 2, which display the number of GCMs that result in an increase in mean annual runoff when the GCM projections are applied to Mac-PDM.09. An important conclusion is that not all simulations are in agreement on whether location-specific runoff will increase or decrease with climate change. The east coast of South America is one location that demonstrates this. Here, the ensemble mean shows that runoff decreases by around 15–20% for 2°C global-mean warming. However, 7–10 models in the same region show an increase in mean annual runoff here. There are subtle differences between the levels of agreement at different degrees of prescribed global-mean temperature change. For instance, the number of GCMs that show an increase in runoff goes up slightly as global-mean temperature increases for the east coast of South America, south-eastern
Australia and central USA. Although there is considerable uncertainty in both the magnitude and the sign of regional runoff changes across different GCMs with climate change, importantly, there is much less uncertainty with regions that experience large runoff increases (e.g. the high northern latitudes and Central Asia) and large runoff decreases (e.g. the Mediterranean).

To further explore how uncertainties associated with climate model structure propagate through to impacts on river runoff, we investigated the changes in the seasonal cycle of runoff for three catchments: the Liard (Canada), the Okavango (southwest Africa) and the Yangtze (China). The Mac-PDM.09 $0.5 \times 0.5 \text{deg}$ grid cells located within these catchments are presented in figure 3. We selected the Liard in the interests of exploring how climate change might affect runoff in a catchment where the runoff regime is predominantly snow-melt dominated. Also, the ensemble mean projections of runoff change suggest that this is a catchment where mean annual runoff increases with climate change. In contrast, we selected the Okavango because the ensemble mean suggests a decrease in runoff here. These two catchments are relatively small so we also investigated changes in a large catchment, the Yangtze.

Figure 3 presents the seasonal cycle of weighted-area-mean runoff for each catchment for a 30-year period when Mac-PDM.09 is forced with observed climate data (1961–1990; solid black line) and for four increases in global-mean temperature (coloured lines according to the GCM). Similar to the disagreement in the sign of runoff change that is presented in figure 2 for mean annual runoff, there is generally no unanimous agreement in the sign of change for the seasonal cycle. For instance, with the Yangtze, around half the GCMs present an increase in runoff with climate change, whereas the other half present decreases. The magnitudes of the differences increase with climate warming. Even at 6°C prescribed warming, some GCMs present very little runoff change relative to present for the Yangtze, e.g. the MPI ECHAM5 GCM, whereas other models present very large changes, e.g. GISS ER and GISS EH; the large runoff changes with these two GCMs are not surprising, however, given that they were developed at the same institute. There is more agreement between GCMs for the Liard for the months January–April, where all GCMs present increases in runoff. This is associated with increased snow-melt-driven runoff as the climate warms. Notably, there are large increases in April runoff relative to present, which is indicative of the onset of snow thaw and melt being brought forward in the year due to warming. However, there is less agreement between GCMs for the months May–September.

Figure 3 demonstrates that there is also uncertainty across GCMs in the timing of the month of peak runoff. For example, with the Yangtze, the peak runoff occurs in June. With climate change, the majority of GCMs show that, although the magnitude of peak runoff changes, it continues to occur in June. However, some GCMs show peak runoff occurring 2–3 months later (e.g. GFDL CM2.0 and GISS EH). Such seasonal changes have implications for crop production and dam management, so it is important that the uncertainties in these changes are demonstrated to aid the decision-making process.

The changes in runoff under climate change scenarios presented so far will have an impact on global water resource stresses. Figure 4a–c displays the projected number of people (expressed as a percentage of the total global population) who experience increases in water resources stress for different amounts of global-mean
Figure 3. Top row shows the ensemble mean percentage change from present in mean annual runoff for a 2°C increase in global-mean temperature, with the bounds of the Liard (left), Okavango (middle) and Yangtze (right) catchments overlaid (thick black lines); thinner black lines show political borders. Also shown is the monthly mean runoff (millimetre; vertical axis) for each catchment for four increases in global-mean temperature when Mac-PDM.09 is forced with climate data from the 21 GCMs (coloured lines) and present-day observations (solid black line), respectively.
Global hydrology modelling uncertainty

Figure 4. Percentage of the global population that experiences increases in water resources stress assuming nine different increases in global-mean temperature from present, with population following the A1B scenario for the year 2080 with: (a) UKMO HadCM3 GCM only, (b) seven GCMs and (c) all 21 GCMs. (d) Median impacts across the 21 GCMs with the inter-GCM range for A1B population. (e) Same as (d) but also includes the A2 population scenario. (f) Same as (e) but also with the B2 population scenario.

Figure 4 displays a global increase in water resources stress, but some regions of the globe will experience a decrease in water resources stress with climate change (e.g. Arnell 2004). For the purposes of demonstrating the application of HTC to global hydrology modelling, we present only the increase in global water resources stress here. Figure 4 highlights the importance of considering climate model structural uncertainty. For instance, figure 4a shows that, if climate change projections from only a single GCM are applied (UKMO HADCM3), which corresponds to the first priority set of simulations in table 1, the number of people who experience increases in water stress is 10 per cent of
the global population for 6°C warming. If we consider seven GCMs (figure 4b), which correspond to the second and third priority sets of simulations presented in table 1, we see that this value now ranges between around 6 and 20 per cent. This range increases further when all 21 GCMs are considered in figure 4c. Although the different GCMs result in different projections of water stress, they are all in agreement that steep increases in water stress occur up to 3.0°C, after which the rate of increase with global-mean temperature flattens. A more useful measure is perhaps the median impact across the 21 GCMs. This is presented in figure 4d, with an ‘uncertainty range’ shaded in grey, which reflects the maximum and minimum possible impacts across all 21 GCMs. This means that, for a given population change scenario (A1B in the year 2080) and global-mean temperature increase of 6°C, the central estimate based on results from 21 GCMs is around 12 per cent of the world population, although it is also possible that this value could be around 6 or 29 per cent. Nevertheless, despite uncertainty, the simulations suggest that the global population exposed to water resources stress could increase by at least 6 per cent for a global-mean temperature increase of 6°C from present. In the present analysis, it is assumed that all the GCMs are equally credible (although they are not completely independent). This is a reasonable assumption because of the challenge in defining appropriate measures of relative performance (Gleckler et al. 2008), but this does require further investigation.

Figure 4e,f explores the role of an additional source of uncertainty, population change. The impacts presented in figure 4a–d all assume population changes in accordance with the A1B scenario. However, population might change differently from A1B, so we used the 189 runoff simulations from Mac-PDM.09 to compute global water stress with two other population scenarios, SRES A2 and B2 (Nakićenović & Swart 2000). The median values are presented in figure 4e,f. The uncertainty range increases when the A2 population scenario is included (figure 4e), but median impacts differ little from the A1B scenario. Inclusion of the B2 scenario (figure 4f) has no effect on the uncertainty range. This demonstrates that uncertainty due to the climate model structure is much greater than that due to population change.

8. Future directions

The results here present the first application of the new ClimGen/Mac-PDM.09 integrated model on the University of Reading Campus Grid. Following the initial success, the integrated model is now being run on the Campus Grid to produce a number of other large ensembles, to exploit the efficiency of the integrated model and gain further value from its development. A further set of simulations have been completed for the NERC QUEST-GSI project comprising a 204-member simulation; the Campus Grid has effectively reduced total simulation time from 816 h (running on a single PC) to 10 h here. Another set of simulations recently completed consists of a 420-member ensemble (21 GCMs, four future time periods and five scenarios) for the AVOID research programme (www.avoid.uk.net), which uses climate change scenarios included in the Committee on Climate Change (2008) report. These simulations took 24 h to complete on the Campus Grid, whereas on a single PC, they would have taken 70 days of continuous computer run-time to complete.
The pattern-scaling approach to calculating climate change projections employed here relies on the assumption that the pattern of climate change (encompassing the geographical, seasonal and multi-variable structure) simulated by GCMs is relatively constant for a given GCM. The direct use of GCM data instead of patterns would avoid this assumption, but would involve several orders of magnitude more input data. GCM data repositories such as the WCRP CMIP3 multi-model dataset repository (www-pcmdi.llnl.gov) at the Lawrence Livermore National Laboratory (LLNL) in California hold much of these data, but it would be impracticable to transfer all these data. It would be far better for running impact models to be able to access the GCM data repositories directly, transferring a subset of the repository data files on demand for temporary local storage on disk or in memory during model execution (Froude 2008).

Additional computational challenges would include that (i) the amount of input data transfer would be larger, even if only parts of input files were transferred on demand, and (ii) SSH access to large, shared repositories is unlikely to be given to individual users because it would require each user to have their own shell account and SSHFS would not therefore be suitable for accessing such remote repositories. One substitute for SSHFS could be the Open-source Project for a Network Data Access Protocol (OPENDAP—http://www.opendap.org), which has been used successfully in the past for modelling projects involving the Reading Campus Grid (Froude 2008). For this to be used to access GCM data repositories, there are three conditions that would have to be met: (i) repository owners would have to be running an OPENDAP server (as LLNL already do, see http://www2-pcmdi.llnl.gov/esg_data_portal/dapserver), (ii) data would have to be stored in the NETCDF format (http://www.unidata.ucar.edu/software/netcdf/), and (iii) the GHM source code would have to be modified to read data from NETCDF files. OPENDAP would not address the problem of model output accumulating in the Campus Grid’s small storage area, and therefore the use of SSHFS with these models is likely to remain necessary.

Another possible approach to remote repository access is to use PARROT (http://www.cse.nd.edu/~ccl/software/parrot), another client utility that, like SSHFS, allows unmodified programs to access remote data through the local file system interface. PARROT has been used on the CamGrid Condor pool at the University of Cambridge (http://www.escience.cam.ac.uk/projects/camgrid/parrot.html). PARROT could be used to access remote data repositories that are not available via OPENDAP and is likely to be more suitable than SSHFS because it is compatible with a wide range of remote I/O protocols including HTTP, FTP and GRIDFTP, which are more likely than SSH to be available to remote repository users.

9. Conclusions

The application of e-Science, and specifically HTC on the University of Reading Campus Grid, has enabled a 189-member simulation of changes in global runoff and water resource stresses under climate change scenarios. This required the integration of two separate models that were primarily designed to be run on a single PC: (i) the Mac-PDM.09 GHM and (ii) the ClimGen climate scenario generator. Resources required for the initial development of

Phil. Trans. R. Soc. A (2010)
the ClimGen/Mac-PDM.09 integrated model and the adaptation of the Campus Grid have been far exceeded by the benefits in improved ensemble-creation time realized by running the model on the Campus Grid. Importantly, simulation time has been reduced from over 750h to 9h, but only when exclusive access to a large proportion of the Grid’s storage resources can be guaranteed. A more practical data management solution involving the modelling team’s own data storage resources was developed for general use, and this allows 189 model runs to be completed on the Campus Grid in less than 24h.

The exploration of changes in global runoff and water resources associated with climate change patterns from 21 GCMs has allowed for a novel and systematic treatment of climate model structural uncertainty on climate change impacts. Although several studies have applied multiple GCM projections at the local scale, none has so far applied this over the entire global domain. The results presented here have demonstrated how important it is to do this and highlighted that it is dangerous to consider only climate projections from one GCM. This is because both the magnitude and the sign of the projected impact can be different between GCMs. Nevertheless, there is a high degree of certainty in the large runoff changes projected for the high northern latitudes, Central Asia and the Mediterranean. The global population simulated to experience an increase in water stress in a 6°C warmer world than present ranged between around 6 and 29 per cent, assuming the A1B population change scenario for the 2080s. Uncertainty due to population change was shown to be considerably smaller than the effect of climate model structural uncertainty.

Although recent developments in e-Science have enabled the production of large ensemble model outputs, it is believed by some that it has also disguised critical questions about what models can usefully offer and how the outputs are used by decision-makers and politicians (Kerr 2009). With this in mind, further work will examine the effect of different ways of combining and comparing results from multiple GCMs. Further work will also exploit the Campus Grid to explore the relative role of GHM uncertainty along with climate model and population uncertainties. Nevertheless, the results here go some way to demonstrating the range of climate change impacts based on the set of IPCC climate model projections available to the impact modelling community. Importantly, this would not have been possible without collaboration between the e-Science and climate change impact modelling communities.

This work was jointly supported by a grant from the Natural Environment Research Council (NERC), under the QUEST programme (grant no. NE/E001890/1), and the e-Research South Consortium (www.eresearchsouth.ac.uk; EPSRC Platform Grant reference EP/F05811X/1). The climate model projections are taken from the WCRP CMIP3 dataset (http://www-pcmdi.llnl.gov/ipcc/about_ipcc.php). ClimGen was developed by Tim Osborn at the Climatic Research Unit (CRU) at the University of East Anglia (UEA), UK. The authors would like to thank David Spence at University of Reading IT Services and the entire Campus Grid team for their support throughout this project. The authors thank Dr Elaine Yolande Williams at the Department of Ergonomics, Loughborough University, and Ms Laura Bennetto at the Design and Print Studio, University of Reading, for their help in formatting the figure images for publication.

References


FAO 1995 *Digital soil map of the world*. CD-ROM. Rome, Italy: FAO.


