

Report of the workshop on “weighting, credibility and reliability” held during the ENSEMBLES General Assembly in Prague, Czech Republic, on Thursday 15 November 2007

Albert Klein Tank (AKT, KNMI), Geert Lenderink (GL, KNMI), Clare Goodess (CG, UEA)

Scope:

Issues related to model weighting, credibility and reliability have been discussed for 90 minutes by about 50 participants. The workshop has been truly cross-cutting across ENSEMBLES expertise and interests with involvement of climate change and s2d modellers, as well as impact groups.

Some specialists could not attend the workshop, but have sent in their comments afterwards. We have added these follow-up comments (in particular from Lenny Smith (LS, London School of Economics) and Hayley Fowler (HF, Newcastle University)), because they are very relevant even though others were unable to respond again afterwards.

Agenda:

In a poll of opinions all participants were first asked to give their immediate response on the 10 selected questions (Q1 – Q10) by showing green or red cards. The outcome of these opinions has subsequently been discussed in three breakout groups gathering around posters with background information (see http://ensembles-eu.metoffice.com/meetings/GA4_Prague_2007/Weightingworkshop_thurs_pm.pdf). The questions were based on recent papers, including ICTP/DMI (2006), Dessai and Hulme (2004), Collins (2007), Fowler et al. (2007), New (2007) and Lenderink et al. (2007). For the sake of discussion, the chosen formulation for the questions was rather black-and-white and probably too confrontational for some. Clearly, there hasn't been enough time for all issues to be discussed in detail.

Questions:

Q1. When do you think a climate model is credible? A climate model is credible...

70%: ...if the climate physics/processes are well represented.

30%: ...if the simulated control climate for the important variables is close to the observations.

The general belief was that we always need verification of models against observations, also because this is convincing for users of model projections. This evaluation should be done at a spatial scale which is in agreement with the scale of the impact model under consideration. However, for future model improvement evaluation of the physics/processes is more important, because the models may be right for the wrong reason when errors cancel out. If the physics are not right, can we trust the future predictions? This is also the reason for the higher score of 70%.

When using information about model credibility for weighting purposes, it was advocated to follow a clear strategy and communicate this strategy. Some suggested a two step approach, thereby combining the different model evaluation approaches. First, make a

selection of models based on the representation of the physics. Next, apply weights proportional to how well the remaining models represent the control climate when judged against observations. This requires different metrics for each step: measures for the representation of the physics/processes in addition to more traditional metrics for model performance against observations. It is also possible to make these two steps scale dependent. GCM measures of performance can then be based on the physics (including how well the large scale circulation patterns are simulated), whereas RCMs can be evaluated by analyzing how well the simulated control climate (using ERA40 boundaries) resembles the quality controlled, high resolution gridded observational datasets for Europe.

Follow-up:

LS argues that an ensemble of climate models is most credible when its member-models can each shadow the observations in hand, and their trajectories (in extrapolation) do not do anything physically silly. As models improve toward decision-support relevance, the focus needs to change to “can the model shadow the relevant phenomena in the past” with sufficient accuracy for us to consider it fit for purpose? This is a much higher hurdle and, of course, depends on the purpose. Still these are only necessary conditions, not sufficient (we hope we have not left out anything important).

HF thinks that there needs to be some assessment of both. As a first test of credibility the climate physics/processes need to be well represented – i.e. if for example the ocean or atmospheric circulation dynamics in the models are wrong then we may not expect them to simulate ‘climate’ well anywhere. However, this is just a first step. The next step might be do the models reproduce the mean global temperature trend, the equator-pole temperature gradient and land-sea contrasts. Believe it or not many of the models in climateprediction.net do not fulfil these basic criteria. The third step might be can we reproduce the important climatic and oceanic modes of circulation: ENSO, NAO, Indian Monsoon etc. This is still at a global scale but it is important to look at the large spatial modes. The fourth stage might be does the model simulate regional processes well – this may be substantiated through comparison of for example regional temperature or rainfall means or, and this is important, the spatial variation in this variables across the region in question. Finally, specific regard must be made to which particular climate variables are important to the impact study under question and whether these can be well represented by the climate model. However, a model becomes credible after stage three, but it still may not be useful for impact studies.

Q2. Weights...

40%: ...should be model specific (i.e. dependent on overall model behaviour).

60%: ...should be application specific (i.e. different weights are assigned for different impacts assessment applications).

The spatial scale is of primary importance when applying weights. Applying general weighting schemes to for instance RCMs based on their behaviour over the whole domain is probably most objective, but not ideal if users are interested in applications at particular sites or areas. What is considered to be the best model overall may not be very good for the variable and region of interest. An impact and parameter/variable dependent weighting scheme is then preferable. This explains why the option “weights should be application specific” scores slightly higher (60%). On the other hand, varying the weights

for each impact model may lead to problems with consistency across impact models in an integrated assessment.

Within ENSEMBLES we should be very careful about model specific weights because our models are far from independent from each other. This is the case in particular for RCMs. Many RCM implementations stem from the same breed. Most simulations also use the same GCM boundaries. In addition to credibility of individual models, similarity between models plays an important role in weighting. If there is an overrepresentation of models that share the same or similar physics, there is a danger that these models get too large weights in the final outcome. A weighting scheme should compensate for this effect.

Avoiding weighting altogether was not seen as the solution. Popular methodologies of weighting (using Bayesian or simpler methods) lead to loss of internal consistency between the variables that are needed in probabilistic predictions to feed impact models. Only multivariate approaches avoid this shortcoming. Generating PDFs involves (perhaps too) many assumptions about the distributions, but in the end probabilistic approaches are needed for many applications (not all).

Follow-up:

LS argues that if the impact model requires a trajectory (time series) as input, then the weighting has to be done after passing that trajectory through the impact model. There are s2d users who do this everyday with medium range ensembles! You have to weight in impact space when nonlinearity is involved; the average value of a nonlinear function of x is not equal to the nonlinear function of the average value of x . Using weights does not imply any loss of internal consistency between the variables if the process is done correctly.

HF actually thinks that the answer to this question depends on the application. For global evaluations it makes sense that the weights should be model specific. Once you get down to the regional scale, weighting should probably be impact specific and based on a combination of the model credibility criteria as mentioned in the answer to Q1. For impact studies we should be looking for a model specific weighting but over a specific region. This should be based not only on those climate variables of most importance to the impact application but more general statistics. For example, if we are interested in flooding we should be interested in models that well reproduce rainfall extremes. However, we should also include in our weighting scheme some reflection of how well models reproduce mean and variability of rainfall as well as extremes and perhaps some measure of temperature. For some applications weighting does not seem to make much difference. However, it is important that we do assess models on their credibility and then weight them – this is one of the biggest questions in our science currently. HF is not sure that Bayesian methods are the answer however.

Q3. Weighting of the GCMs and RCMs should be done...

50%: ...in combination for GCM-RCM pairs.

50%: ...separately.

The discussion from Q1 and Q2 is partly repeated, focussing also on the measures that should be applied. This includes GCM weights based on physical metrics, global climate sensitivity, circulation, mean temperatures and RCM weights based on ERA40

simulations judging the added value (their downscaling ability) on the smaller scales. One can also argue that the performance of the prediction system should be measured by analyzing the output of the impact models rather than the climate models. Does this mean we should weight for GCM-RCM-impact model pairs?

The general confusion on this question is best expressed by the 50-50 score. Again, a clear strategy is urged for. The ENSEMBLES approach should not be decided within a single RT, but rather result from discussions among different RTs.

Follow-up:

LS argues that unless the process is linear, the weighting has to be done based on the output of the combined process.

HF finds this a difficult question. For many reasons it would be very difficult to separate the influence of the GCM from that of the downscaling RCM and therefore HF would suggest that GCM-RCM pairs are weighted. However, if there is just a suite of GCMs then we should obviously just weight separately. HF does not believe that the performance of the prediction system should be measured by analyzing the output of the impacts models rather than the climate models. Bringing the impacts model in at this stage just provides extra complexity when what we are actually interested in is which climate models will give the best predictions over a region. However, there may be a need for a separate step in impacts studies where the ability of the impacts model is assessed against other types of models (in terms of both structure and parameterisation) – there are a few studies that have done this in a simple way already.

Q4. Many aspects of weighting and reliability assessment in seasonal and interannual prediction can be used for climate change prediction.

50%: Yes, the experience from the short time scales improves climate change prediction.

50%: No, different approaches are required.

Again 50-50. However, this time this score doesn't seem the result of general confusion, because there are clear and well informed opinions from both s2d and climate change modellers in the group.

Some argue that in a seamless prediction approach the evaluation should be similar across the different timescales. Therefore, the experiences gained and lessons learned in seasonal and interannual prediction provide important input for model evaluation and weighting at the longer timescales. However, the reliability of the predictions on the short timescales can be assessed from past cases, whereas for the long time scales no such verification of the ensembles prediction system is possible. Moreover, s2d prediction is an initial value problem, whereas climate change prediction is a boundary value problem. Different questions are asked for the two time scales. Important feedbacks related to processes do not play a major role in s2d prediction, whereas they are decisive on climate change timescales.

Follow-up:

LS finds it silly to ignore the relative value of a model in reproducing seasonal phenomena of empirical relevance or user interest. At the same time, extrapolation into new climate parameters could change the relative value of models. That said, it seems unlikely that a model that was unable to reproduce the phenomena of interest better than

climatology in the current climate would be considered "fit for purpose" answering user questions in an altered climate if that purpose meant getting realistic variability which the model cannot do in the current climate. Still that model might be of value at larger space and time scales in both instances. Climate change is both an initial and boundary value problem (e.g. through the oceans on time scales of a hundred years). Regardless: the impacts depend on the details of the weather trajectory, we must sample this if we are to be empirically or decision-support relevant.

Q5. Do we sample the uncertainty sufficiently to derive useful probabilistic climate change projections? The ENSEMBLES GCM-RCM matrix is...

10%: ...sufficiently filled.

90%: ...inadequately filled.

It is generally realized that the ideal situation would be a far larger number of RCM simulations filling in the GCM-RCM matrix. However, it was also argued that, given the limited computer resources and the therefore limited number of RCM integrations, the filling in of the matrix has been done optimally. That is, the filling of the matrix has been done in such a way that we could use statistical methods to fill in results for those runs that have not been performed. This implies that a sufficient number of RCM runs for a few GCMs has to be performed in order to assess the downscaling uncertainty.

Besides the issue of the GCM-RCM matrix, it is realized that with the choice of the GCMs we do not sample all uncertainties sufficiently. As such, the use of one emission scenario, neglect of carbon cycle feedback, and under-sampling of the uncertainty in the climate sensitivity could be mentioned. As a result, the uncertainty in the global temperature response is under-sampled. Therefore, statistical techniques, like pattern scaling, are needed to broaden the uncertainty range.

Follow-up:

LS: define "probabilistic". If you mean "decision relevant probability forecasts" then almost certainly not. If you mean an "ensemble of potential pathways" of more value than "one run per forcing scenario" then certainly yes. The details of the application matter here. The matrix is certainly incomplete, but that does not make it without value! This relates also to questions Q9, Q10.

HF thinks that the answer to this is "inadequately filled". But how useful is this question? What would be more useful is to determine where the largest uncertainties in the system lie and to concentrate on sampling them better. At the moment we have a rather random set of models (dependent on history). Analysing perturbed physics ensembles may provide a better grasp of which parameter ranges are valid and therefore a better process understanding. ENSEMBLES provides a range of models – what is more 'realistic' cannot be defined. However, methods such as pattern scaling do not provide adequate solutions. If we can use these kind of methods successfully to provide information for different time periods, emissions scenarios, or (as we are doing in the UK, as part of UKCIP08, climate model parameterisations!) then why do we need to run the complex models that take up so much computation time? Pattern scaling is not the solution – however, distributed computing such as demonstrated in CP.net may well be.

Q6. Do we need RCMs as downscaling tool, or are statistical methods sufficient?

100%: Yes, we need RCMs.

0%: No, statistical methods can do the job.

The usefulness of RCMs is not questioned. However, most people also felt that statistical downscaling is needed. The complementary question “Do we need statistical downscaling...?” is also thought to be true. It is generally realized that we need statistical tools to fill in results for those RCM runs in the matrix that have not been performed. It is not clear whether statistical downscaling should be limited to additional filling in of the matrix or whether statistical downscaling should be applied as a separate method besides dynamical downscaling to produce regional scenarios. In the latter case both RCMs and statistical downscaling will have their own downscaling uncertainty. This may lead to large uncertainty bands in the downscaled climate. Do we want this, and if not, how can we constrain the outcome to what we think is a more realistic range? In the former case (additional filling in) we have to think about how to constrain the statistical downscaling in a way that is consistent with the RCM results. We can test the ability of the statistical downscaling to reproduce the RCM results (offering a possibility to “validate” the system).

Follow up:

GL: The KNMI employs a scaling technique using not only global temperature, as in pattern scaling, but also strength of the westerly circulation, to do the statistical filling of the matrix (Lenderink et al. 2007). The downscaling is constrained by the season mean changes in precipitation and temperature in an ensemble of GCM results.

LS: Do we have any empirical basis to believe one approach will extrapolate better than the other? Dynamical and statistical methods are complementary, and the best approach might well include a bit of each.

HF’s opinion on this has changed over the last 5 years or so as RCMs seem to have become more sophisticated. It is important to keep on developing RCMs. For local scale impacts these are the only way of assessing potential dynamic changes in the environment as statistical methods assume stability between large scale atmospheric processes and local scale climate. A combination of dynamical methods (to provide local scale changes and changes to extremes which cannot be well simulated by GCMs) and then their application through statistical methods (such as the weather generator approach developed for the UK – EARWIG) is probably the way forward here. Unfortunately, in many cases, RCM outputs cannot be used directly as they are biased. Therefore taking the changes from RCMs and applying this to simulation models built using observed data may be a good way forward.

Q7. Verification of the ENSEMBLES prediction system...

80%: ...can be done.

20%: ...is an impossible task.

This is a surprising outcome. There is no question that ENSEMBLES produces a large amount of very useful information about climate change. However, it is unlikely to produce a system that will be completely consistent with our knowledge about climate change. A benchmark could be to test the systems against IPCC AR4. It was argued that

the knowledge has already been advanced compared to the IPCC AR4, so that this may not be a valid comparison.

Most participants would rather reformulate the question to “The verification of the system **should** be done”. We should try to find ways to “evaluate” the performance of the system. It was proposed that we could pursue a “perfect” model approach; that is, assume that one model is the real climate, build a system based on the remaining models, and test whether the system is consistent with the perfect model. Another way of evaluating such a system is to investigate the robustness of the system to assumptions.

Unlike the season forecasting system where there are methods to verify the system in a statistical sense, we have no means of evaluating the system since we only have one realization. This could mean that we have to redefine our interpretation of uncertainty.

Another related important issue is that we should try to define what we mean by probabilistic climate change projections. This relates to question Q9 and Q10. Is what we mean with probabilistic climate change projections useful for society?

Follow-up:

According to LS, verification of a probabilistic s2d forecast system for two weeks ahead has been done, and value established (for decision support and empirically against climatology). The very idea of "verification" (better: evaluation) of an extrapolation differs (unless one is willing to wait and see). Model diversity is a blessing when we have the observational information to determine which models are of greater value and exploit (weight?) them. In climate change, however, model diversity (after sampling ICs) implies that the details matter, and until we resolve those details (by physical argument) we have ambiguity on the space and time scales where different model-classes give different answers (in distribution). This ambiguity is not described by a probability-like-uncertainty. We can always form a subjective probability, but then whose subjective belief is it? Ambiguity means there is a good chance what will happen is not even in the range of things our models offer us as possibilities.

Q8. Probabilistic climate change projections should...

95%: ...include natural variability.

5%: ...exclude natural variability.

A large majority voted for including natural variability. In hindsight this could be the consequence of the way the question was formulated. It is generally realized that changes in the variability are an important factor of climate change, perhaps or even likely, more important than changes in the means. This could explain the large majority for “include” natural variability.

Indeed, a large part of the discussion did focus on what we meant by natural variability – e.g., decadal variability, the movement around the mean, the noise in the system – and the extent to which it can be predicted/quantified. It was agreed that it is difficult to know how much of natural variability the observations/models are capturing since we only have one realisation of the past. It was suggested that we should exclude the ‘unpredictable’ part of natural variability and focus on the part which is likely to be affected by changes in forcing – though how this can be done was not addressed. One reason given for excluding natural variability from the projections was the argument that we should avoid

considering ‘futures that may never occur’. Others argued that ideally users should be presented with information about changes due to emissions scenarios together with separate, accompanying estimates of natural variability.

Follow up:

GL: If we compare a 30-year control period with a 30-year future period, and derive climate change projections from these two periods, we do not only sample the climate change signal, but some of the changes may also be due to natural variability (and therefore are due to pure chance). The purpose of the question was whether we should filter this signal due to natural variability out or whether it should be retained? Note that the change in two 30-y periods due to natural variability may not always be relevant to the impact models. For example, for river discharges often maximum discharges with return periods of 100-1000 years have to be estimated, in which case natural variability based on 30-y periods strongly obscures the climate change signal (see e.g. Lenderink et al. HESS, 11(3), 1145-1159, 2007). Note also that statistical pooling techniques or taking ensemble averages all have impact on the change signal due to natural variability only. On the other hand, it is important to convey the message that not all changes are due to anthropogenic climate change.

From a mitigation point of view, it may be better to only give climate change signal, since reduction in emissions are motivated by our impact on the climate system (and therefore the climate change signal). Irrespective of the choice, we should communicate very clearly what is done. Or coming back to the earlier remark, we should clearly define what we mean with probabilistic climate change projections.

LS remarks that the separation between climate response and natural variability is artificial, and cannot be done with a non-linear system as the climate system. Therefore climate change projections should include both.

HF thinks probabilistic climate change projections definitely need to include natural variability. For scientists and non-scientists it is interesting to see separate PDFs of natural variability and climate change. However, for design purposes – to enable adaptation to climate changes – we need to know not only the climate change but the envelope of natural variability. Very wide PDFs of wind extremes are appropriate as climate models are very bad at simulating wind never mind the extremes of wind. Local scale rainfall should also have a wide PDF as it will do naturally. What is important to communicate is what is the magnitude of the change that we can expect on top of the natural variability that we currently experience? Or will natural variability itself perhaps become more variable as the world warms?

Q9. Impact models need and/or can handle...

85%: ...probabilistic climate change projections.

15%: ...a limited set of indicative scenarios describing possible future climate states.

According to the discussion, both answers are credible and not necessarily mutually exclusive. PDFs are important for understanding/interpreting deterministic scenarios within a wider context of uncertainty, i.e., they can provide a range within which to consider the single scenarios. The appropriate or feasible approach depends in part on the complexity of the impacts model and the computing power available. We also need a better understanding of the sensitivity of impacts models and the associated

uncertainties, e.g., it would be good to have some perturbed parameter ensembles based on impacts models. Some users will be able to benefit from using PDFs, while for others using a few scenarios may be the most practical, and still useful, option.

It was noted that some of the results presented earlier in the meeting (e.g., the Hadley Centre perturbed physics ensemble (PPE) extended to carbon cycle feedbacks) indicate that we may be in danger of underestimating the 'top end' of the warming range at the end of the century. This led to a broader discussion of whether we can actually produce probabilistic information using a combination of GCMs, RCMs, and statistical downscaling. E.g., pattern scaling may not be appropriate at the 'high-end' – and downscaling stabilisation scenarios and carbon cycle PPEs are issues which have not been addressed by ENSEMBLES. Ideally consistent approaches should be taken across the different applications sectors – but there is, for example, the consideration of whether any weights used should be climate model or application specific (see Q2).

It was also noted that the scenario producers would benefit from more feedback from the impacts users, e.g., RT6 had found that three of the s2d models were giving significant skill. It would be good to collate and feedback such information to the modelling groups. Perhaps a wiki discussion forum could be used to provide such feedback and to ensure ongoing dialogue within an end-to-end approach.

It was suggested that any sort of probabilistic information was useful – although it is important to present the underlying assumptions. Users should be reminded that ENSEMBLES is producing probabilistic projections conditional on specific emissions scenarios (e.g., A1B) – and this conditionality assumption is very important from the socio-economic impacts/adaptation perspective.

Follow-up:

According to LS, it is not easy to do climate science, and impacts analysis is one step harder. What we feed them should be based on what we know; and limited by what we know we do not know well enough to be useful/relevant/"fit for purpose". Almost all impacts and user decisions require trajectories in time. Indicative scenarios can be of value if we do not think we can provide robust decision-relevant probabilities (trajectories with a relevant distribution). A selection of possible pathways, presented as incomplete but useful, can have lasting value and allow good expectation management.

According to HF, impact models can handle probabilistic climate change scenarios, but what we have at the moment are not fully probabilistic scenarios. They are a set of climate model runs based on history – i.e. an ensemble of opportunity. At the moment we are only looking really at an indicative set of scenarios describing possible future states. We need to also incorporate the uncertainties in impact modelling into this probabilistic approach. Some impacts groups (in particular in academia) are able to post-process the large amounts of data needed to examine probabilistic climate change scenarios. However, industrial and consultancy users would not be able to use this and so instead, they typically use an extremely limited set of scenarios. We need to think about how we, the scientists, can produce valuable information and tools to be used by non-scientist users where they do not need a detailed understanding of what is going on. Software tools can be produced that are easy to use and allow the investigation of probabilistic scenarios in a limited way with a black-box approach; users do not need to know all the science behind such tools despite what scientists might think. However, even with such sophisticated tools, users will inevitably pick only a few scenarios to concentrate their

efforts upon. We need to provide more guidance so that these better cover the range of uncertainties in response.

Q10. Society needs...

20%: ...probabilistic climate change projections as a basis for rational decision making.

80%: ...a limited set of indicative scenarios, each describing possible future climate states.

It was thought that many of the issues discussed under Q9 were relevant to this question. In the particular context of 'society' and stakeholders (rather than applications users who were the focus of Q9), it was thought that training and education were particularly important. 'Translation' may be needed, e.g., what are PDFs? The focus should be on those variables that people need. There was, however, a view that stakeholders still want 'black and white' answers and may not be ready for probabilistic information.

Follow up:

LS argues that there is a danger in overstating our confidence in probabilistic climate change projections. If we provide meaningless random numbers in a user-required format, why will they pay any attention to us in 10 years when we have relevant numbers?

GL: Another issue is that many processes in society will not be governed by objective cost-loss calculations. As a scientist we may overrate the capacity of society to decide based on objective criteria (if we could provide them). Decisions are made based on a range of constraints and arguments, of which climate change is only one. Is decision making ready for probabilistic climate change projections? or can a few indicative scenario's be sufficient or even better to support decision making? Perhaps the answer depends on the scale or application. It may be possible to constrain the global climate response to a certain level (global temperature rise) with a certain probability. However, local adaptation to increasing precipitation extremes could be another chapter.

HF: We need to identify adaptation options that are robust across the full range of uncertainties. Probabilistic scenarios – should we be able to provide fully probabilistic scenarios across the full range of responses – would be the ideal solution. However, the best that we can do currently is probably to look across as large a range as we can using different methods etc. to define the ranges. I think that weighting may help here though – for decision making at the local scale we should really only use those models which are 'credible' for that particular critical impact. And so we go full circle...!

Conclusion:

The workshop has been successful and stimulating to continue the discussion of this important cross-cutting topic within ENSEMBLES. Time was too short to decide how to bring this further in practice, but we believe it is up to the management board to provide follow-up in the form of cross cutting initiatives. The topics discussed will certainly come back on the agenda of future General Assemblies. In the meantime, further comments on this paper, including perhaps suggestions on rephrasing or better defining the key questions, are welcomed.

References:

Collins, M., 2007. Ensembles and probabilities: a new era in the prediction of climate change. *Phil. Trans. R. Soc. A*, **365**, 1957-1970.

Dessai and Hulme, 2004. Does climate adaptation policy need probabilities? *Climate Policy*, **4**, 107-128.

Fowler, H.J., S. Blenkinsop, C. Tebaldi, 2007. Review: Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. *Int. J. Climatol.*, **27**, 1547-1578.

ICTP/DMI, 2006. ENSEMBLES Deliverable D3.2.1: Definition of measures of reliability based on ability to simulate observed climate in hind-cast mode.

Lenderink, G., A. van Ulden, B. van den Hurk, and F. Keller, 2007. A study on combining global and regional climate model results for generating climate scenarios of temperature and precipitation for the Netherlands. *Clim. Dyn.*, **29**, 157-176.

New, M., (2007). Probabilistic regional and local climate projections: false dawn for impacts assessment and adaptation? TGICA Meeting Report: Papers Regional Expert Meeting. In press.