Climatic Change (2013) 117:1–9
DOI 10.1007/s10584-012-0551-0

COMMENTARY

Testing ensembles of climate change scenarios for "statistical significance"

Hans von Storch · Francis Zwiers

Received: 6 February 2012 / Accepted: 13 July 2012 / Published online: 11 August 2012

Abstract Climate impacts and adaptation research increasingly uses ensembles of regional and local climate change scenarios. To do so, the ensembles are examined to evaluate whether they describe a systematic difference between present states (and impacts) and envisaged future states—and such differences are often characterized as being *s*tatistically significant. This term "significance" is well defined by statistical terminology as the result of a test of a null hypothesis that is applied to samples of observations that are obtained with a defined sampling strategy. However such a statistical null hypothesis may not be a well-posed problem in the context of the evaluation of climate change scenarios. Therefore, the usage of terms such "statistically significant scenario" may be misunderstood in the general discourse about the certainty of projected climate change. We propose to employ instead a simple *descriptive approach* for characterizing the information in an ensemble of scenarios. Physical plausibility in the light of theoretical reasoning often adds robustness to the interpretation of climate change scenarios.

H. von Storch
Institute for Coastal Research, Helmholtz Zentrum Geesthacht-
Zentrum für Material- und Küstenforschung, Geesthacht, Germany
e-mail: hvonstorch@web.de

F. Zwiers (✉)
Pacific Climate Impacts Consortium, University of Victoria,
Victoria, BC, Canada
e-mail: fwzwiers@uvic.ca

1 Introduction

Ensembles of climate change scenarios, say of future seasonal precipitation in Northern Germany, are sometimes described as being *significant* (http://www.climate-service-center.de/031451/index_0031451.html.de; accessed 18 June 2012). In scientific contexts, this term usually refers to an assessment of the likelihood of a given outcome under an assumed set of conditions, with greater significance associated with rarer outcomes under those assumptions.

A key element in determining statistical significance is a *null-hypothesis* H_o, against which statistical rareness is assessed; *significant* results are those that occur in the far tails of the distribution that prevails under H_o. To assess statistical significance under H_o, scenarios are implicitly viewed as realizations of a *r*andom variable or *p*rocess, i.e., as members of a well-defined population of possible events. Decisions regarding H_o are generally made on the basis of a test statistic t that summarizes the available sample of realizations; this statistic is itself a realization of a random variable T, where T represents the population of all possible outcomes of t under repeated sampling. Note that generally the test statistic has been constructed so that it has desirable optimality properties under an assumed set of conditions, although this is not always the case.

How far in the tails, and which directions in the tails, provide information regarding signficance is often a matter of social agreement that may or may not, in part, be explicitly articulated as an *alternative hypothesis* H_a. In climate science, the tail is often given by those events larger than 95 % of possible outcomes under H_o, corresponding to a significance level of 5 % or less. The null-hypothesis is rejected when the outcome t, as summarized by the test statistic on the basis of a collection of scenarios $p_1, p_2, ..., p_n$, is inconsistent with the statistical model proposed under H_o. An underlying assumption that supports the interpretation of t is that the available collection of scenarios $p_1, p_2, ..., p_n$ forms a sample of realizations of a random process P, where the sampling process is understood. That is, the assessment of the rarity of the outcome under H_o depends upon an assumption that the available sample of scenarios is representative of a generating process, or population, P. We will see later when we discuss *ensembles of opportunity* (Tebaldi and Knutti 2007) that this is a non-trivial assumption.

It is understood that decisions concerning the null hypothesis are subject to error with an error rate that corresponds to the selected significance level. Generally, tests of hypothesis only operate with the user-specified error rate characteristics when all assumptions are satisfied. In case of climate change scenarios (for example, simulated differences between future and present precipitation), tests of the null hypothesis that there is no change on average, $H_o : \Delta\mu_P = 0$, versus the alternative hypothesis $H_a : \Delta\mu_P > 0$, are often applied to ensembles of scenarios constructed from multiple climate models. [We employ here one-sided tests for the sake of simplicity.] Rejection of this null hypothesis means that the scenarios tend to be positive, but it does not describe the range of behaviour seen in the scenarios particularly well, since individual ensemble members could show precipitation reductions even if the ensemble mean tends to be positive.

For clarification, we emphasize that we do not speak about the process of generating socio-economic and emission scenarios, but about ensembles derived from such scenarios using climate and climate impact models. The issue here is exclusively

the question, to what extent the assumptions of the statistical methodology used for hypothesis testing are fulfilled or not.

2 Is hypothesis testing of scenarios well-posed?

The response to this question depends very much upon the context within which it is posed.

In the case of a single climate change scenario, produced under a specified forcing scenario with a single climate model of a type that generates its own chaotic internal variability, the question, for example, of whether a projected change in mean precipitation is significantly different from zero, is a well posed question. Such a hypothesis can be tested with a parametric test (von Storch and Zwiers 1999), a non-parametric testing procedure, or by means of a Monte Carlo testing procedure that involves additional sampling from the same climate model, e.g., by generating multiple simulations with that model from randomly selected initial conditions. Regardless of the method, the inference that is drawn in this case applies directly to the climate simulated by the model, is conditional upon the selected forcing scenario, and provides information about the future of the climate to the extent that the model is reliable and future forcing evolves according to the forcing scenario. In this case the null-hypothesis H_o reads: *On average, precipitation does not increase in realizations of model A using scenario B*. Note that this hypothesis can be tested locally at the model grid-point scale, regionally or even globally.

However, our primary concern is with the more general case in which potential future change is described with an ensemble of climate change scenarios that are derived from multiple climate models and perhaps with multiple forcing scenarios. Such ensembles of opportunity encompass variation between ensemble members that is due not just to natural internal variability in the climate system as simulated by climate models, but also model and forcing scenario differences. Here, an appropriate formulation of a null-hypothesis and of the underlying random processes is much more of a challenge than in the previous simpler case.

To explain the concern we will consider two examples, one based on an ensemble of scenarios for future rainfall amount in Northern Germany, and a second based on an ensemble of stream flow projections for the Peace River in British Columbia, Canada. We begin with the rainfall projections as a typical example of projected change in a weather variable. To be specific, we consider n scenarios of possible future climate change at a given location for, say, seasonal summer rainfall amounts $p_1 .... p_n$ (as shown in Fig. 1) . Let us assume that the projected change $p_i > 0$ in m cases and that $p_i < 0$ in $n - m$ cases. In a world without forced change it would be reasonable to expect equal numbers of increases and decreases. That is, we would expect the null hypothesis $H_o : \mathrm{E}(m/n) = 0.5$ (with E() denoting the expectation of a random variable) to be true at all locations. If the p_i are drawn independently (from an imaginary infinite population P of possible scenarios) and if H_o is true, then m has a binomial distribution $\mathcal{B}(m; n, 0.5)$ where the probability for each p_i being positive is 0.5. This is a simple test (the sign test; Conover 1971), but good for explaining concepts. If we make the necessary assumptions, then we can determine an m^* so that $\sum_{k \geq m^*} \mathcal{B}(k; n, 0.5) \leq 5$ %, and we would reject the no-change null-hypothesis against the alternative hypothesis of precipitation decrease at less than or equal to

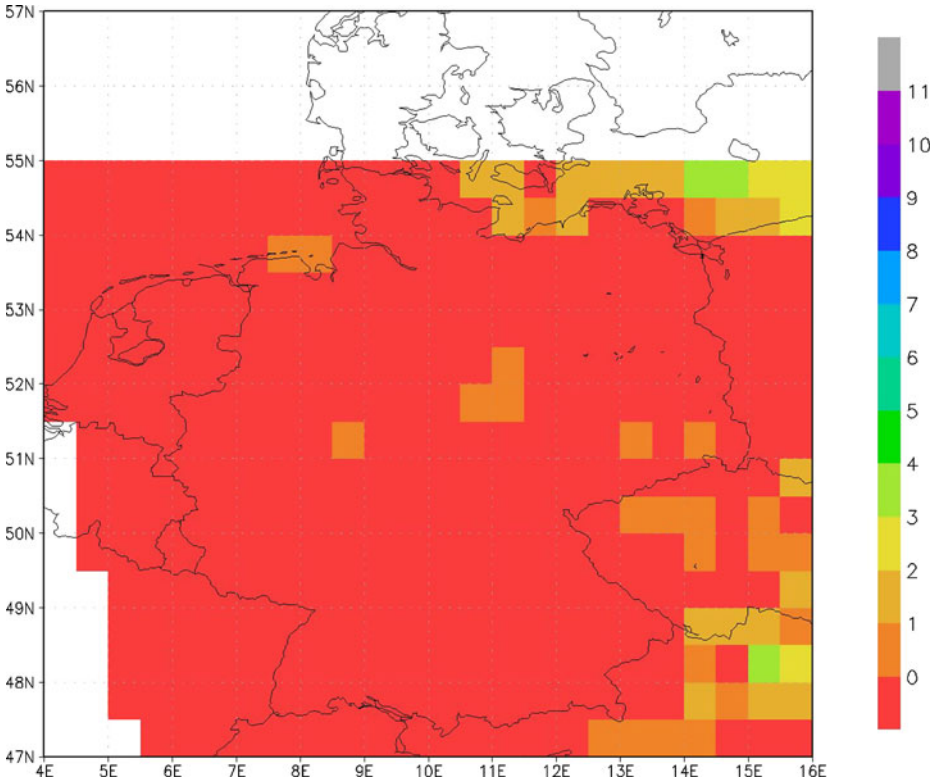

Fig. 1 Number of scenarios exhibiting a positive change in summer rainfall amounts at the time horizon 2071–2100 vs. 1961–1990. The *grid boxes* have a size of about 50 km. *Red* refers to $m = 0$ positive changes (all changes are negative), while *green* represents $m = 3$ or 4 positives changes (out of 11); see color code on the side (Meinke et al. 2010)

the 5 % sigificance level when the actual number m, derived from our specific sample, is less than or equal to m^*. This is because $m \leq m^*$ would be observed in 5 % or fewer cases when the probabilities of a negative and positive outcomes are equal.

Figure 1 shows the number m of positive changes for Germany amongst the eleven scenarios available in the *Norddeutscher Klimaatlas* (Meinke et al. 2010). This ensemble of regional climate scenarios is composed of 11 high resolution regional climate simulations: four simulations from COSMO-CLM operating at ∼20 km spatial resolution and driven with ECHAM5 global simuations (2×A1B, 2×B1); three simulations from REMO at ∼10 km spatial resolution using ECHAM (A1B, A2, B1); four simulations from RCAO/SMHI at ∼50 km spatial resolution using HadAM3H and ECHAM4/OPYC3 global driving simulations (1×A2 and 1×B2 for each global model)). The variable considered is summer rainfall amount on grid boxes of about $(50\ \text{km})^2$. Red boxes (the majority) indicate that $m = 0$ of the 11 cases exhibit an increase; the small green area in the Baltic Sea southwest of Bornholm represents two boxes with $m = 4$, i.e., four scenarios project an increase whereas seven project a decrease in summer precipitation. The uniformity of the pattern is not uncommon in these eleven scenarios; for the winter season, the

map is almost completely reversed, with $m = 11$ almost everywhere. For $n = 11$, the 5 %-critical value of the binomial distribution is $m^* = 2$ (with $\mathcal{B}(2; 11, 0.5) + \mathcal{B}(1; 11, 0.5) + \mathcal{B}(0; 11, 0.5) = 3.3$ %). Thus in the often used parlance, the ensemble of 11 *Norddeutscher Klimaatlas* scenarios is almost everywhere "significant" at less than the 5 % level (with $m \leq 2$); only about three (green) boxes exhibit values of $m = 3$ or $= 4$ that are "not significant".

Should it therefore be said that "the ensemble is significant for Northern Germany"? Unfortunately, such a statistical interpretation is fraught with difficulty. One issue related to the application of the sign test, which is discussed increasingly in the climate science community, is whether the scenario members can be considered to be independent (cf. Woth 2005; Tebaldi and Knutti 2007; Abramowitz 2010). This is nevertheless a minor issue in relation to the broader question—what is the population of potential scenarios that we are referring to? In what sense is the ensemble $p_1, \ldots p_n$ representative? What is P? The difficulty is that the test makes an inference about non-observed cases, namely additional scenarios that could, in principle, be drawn from P by running additional climate models and considering additional forcing scenarios. However, we would be challenged to describe the statistical sampling strategy that led to the available ensemble (Tebaldi and Knutti 2007).

To explore this difficulty further, one might ask whether the inference for Germany applies to:

a) *"All climate scenarios"*—in which case, we would have to determine what is meant by "all". This presumably means climate scenarios produced with all conceivable models, all conceivable emission scenarios, and all conceivable downscaling approaches, plus an understanding of how the available 11 scenarios were selected from that broad population. Moreover, this would need qualification. All emission scenarios that are deemed valid by contemporary economists? Or all followers of a certain school of thought? All climate models? Note that in the IPCC Special Report on Emissions Scenarios (Nakicenovic and Swart 2000) probabilities or relative likelihoods were not assigned to the various scenarios that were described.
   There is likely no way to make an assertion about "all climate scenarios", because that set is simply not definable. Nevertheless, there have been attempts to quantify uncertainty from models, forcing scenarios and downscaling, for example, using complex hierarchical Bayesian models (cf. Murphy et al. 2007, 2009), and serious thought has been given to the basis for the interpretation of statistics calculated from multimodel ensembles (cf. Rougier et al. 2010).
b) *"Climate scenarios based on a specific emissions scenario"*—in which case we would want to make statements that are specific to individual emission scenarios. This is the approach that was used by the Intergovernmental Panel on Climate Change (IPCC) in its 4th Assessment Report (IPCC 2007), in which it assessed a likely range for global projections of future change under each of several different emissions scenarios. The IPCC was also confronted by the question of the interpretation of the available ensemble of models, and thus the assessments of global projections included an aspect of expert judgement that drew on physical understanding as well as the spread of available ensembles of climate change simulations.

c) *"Climate scenarios based on a specific emission scenario and produced with a restricted class of models"*—this is closer to being a tractable problem if the class of models is sufficiently restricted, albeit still a very large problem. For example, one might consider models that share the same code, but in which parameter settings have been varied systematically using a well designed sampling scheme, such as is used in the *climateprediction.net*-project (Allen et al. 2000) (see also Murphy et al. 2004).

d) *"All available climate scenarios"*—in which case this is no longer a problem of statistical inference because the population is completely known. Therefore we can simply state: summer rainfall decreases in most locations in all $n = 11$ cases, with some exceptions at a few locations in $m = 3$ or four cases. No attempt is made to quantify statistically how additional, so far unknown scenarios, would play out. Nevertheless, the lessons learned from the responses to forcing simulated by a limited collection of physically based climate models under a limited range of forcing scenarios, together with our understanding of physics and how changes in the composition of the atmosphere affect the planet's energy balance, should provide considerable insight about the general nature of the responses that can reasonably be expected from other physically based climate models and under other forcing scenarios.

These concerns of interpretation are not restricted only to scenarios of future weather conditions, but also to scenarios of more complex aspects of the future climate, such as future stream flow in large river basins. The possibility that there may be change in future hydrological regimes is of concern both for users of rivers and from an ecological perspective. Figure 2 shows a collection of projections of change in the annual hydrograph (flow regime) of the Peace River in British Columbia (BC) at the Bennett Dam, which generates a large proportion of BC's electricity. These scenarios (see Schnorbus et al. 2011) were produced by first statistically downscaling an ensemble of 23 global climate model simulations for the Peace River basin, and then driving a hydrologic model of the basin with the downscaled projections to obtain simulated stream flow in the Peace River. The 23 global climate model simulations come from eight models that participated in CMIP3 (http://www-pcmdi.llnl.gov/ipcc/about_ipcc.php), and with one exception, three simulations from each model were used, one simulation forced with each of the SRES A2, A1B and B2 emissions scenarios. The figure shows that for this particular river system and gauging site, winter flow is projected to increase in all scenarios, that the timing of peak flow advances in most (but not all) scenarios, and that summer flow is reduced. These are changes that are consistent with warming and winter precipitation increase in a basin that in the present climate, is dominated by winter snow storage (i.e., a *nival* flow regime). Some aspects of these changes would be considered to be "significant" if the sign test were to be applied naively, as discuss above, but as with precipitation change in North Germany, this would be misleading. The projected changes that are shown in Fig. 2 are physically plausible, and do cover a considerable range of uncertainty, but clearly not all uncertainties (e.g., uncertainty that results from the choice of downscaling technique or hydrologic model is not at all represented in Fig. 2—and we are again challenged to understand the sampling process that lead to the selection of these 23 scenarios from a hypothetical population of such scenarios.

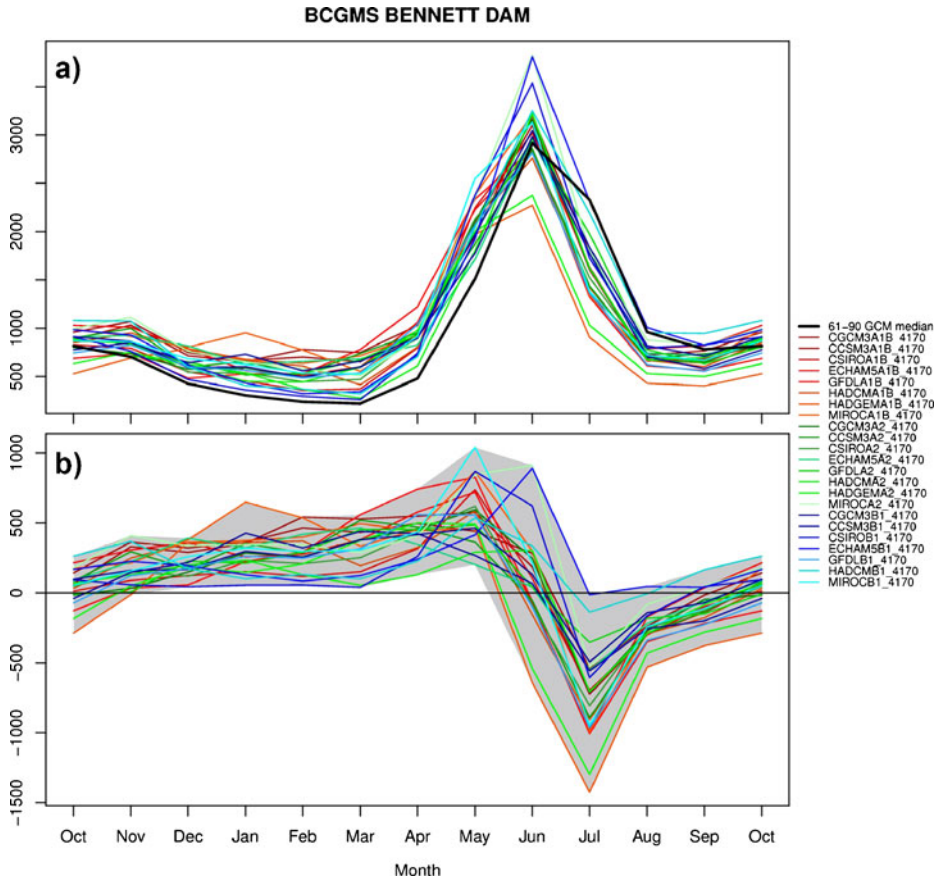

Fig. 2 Median monthly discharge for the Peace River at Bennett Dam showing: **a** historic (1961–1990) and future (2041–2070) discharge, and **b** the 2050s anomaly. Historic discharge (*black line*) is presented as the full ensemble median (23 runs × 30 years) and each future discharge value is the median of each GCM run (1 run × 30 years). Anomalies represent the future monthly median minus the historic ensemble monthly median. From Schnorbus et al. (2011), Fig. 4.4

3 Concluding remarks

The question arises why the term "statistical significance" is used. Presumably, one reason is that such "significance" is often confounded with importance. The lay and statistical usages of "significance" are often mixed in the overall discourse on climate change. For many, this terminology may indicate a certainty when this may not, in fact, be the case. This shrouds the character of scenarios as scripts of possible future change, which should be used by stakeholders to examine possible consequences and possible countermeasures (Schwartz 1991). Among lay-people, this may also contribute to the common blending of the term "projections" with the term "predictions" (Bray and von Storch 2009), in spite of the careful distinctions made in IPCC reports.

Even if statistical testing were completely appropriate, the dependency of the power of statistical tests on the sample size n remains a limitation on interpretation.

In our case with a few localized green boxes, the fact that precipitation increased in $m/n \sim 27$ % of scenarios implied that there was insufficient evidence to rejected the no-change hypothesis H_o even though precipitation decreased in the remaining 73 % of scenarios. On the other hand, if this proportion of the $n = 676{,}339$ *climateprediction.net* HadSM3 simulations (as of 13 December 2011; http://climateprediction.net/) had shown a decrease in the region, rejection would have occurred at a significance level effectively equal to 0 %. However, in both cases, the result would be that 27.3 % of scenarios point to an increase in rainfall amounts, and 72.7 % to a decrease. We would suggest that the strength of discrimination, such as in *recurrence analysis* (von Storch and Zwiers 1988), is an important adjunct to testing for the identification of robust information. This would also help to lay greater emphasis upon the magnitude of the projected change as well as its variability between ensemble members. Depending on a variety of factors, a "significant" change might be so small as to be totally irrelevant from an impacts perspective.

Nevertheless, given the conceptual problems discussed above, it is perhaps best to simply express the state of knowledge in a descriptive manner such as the following: *Using n scenarios constructed with the models A, B, ..., emissions scenarios S_1, S_2, ..., and so on, we find rainfall amounts decrease in most grid boxes for all scenarios, and that in the remaining few grid boxes, they decrease in most (*72.3 *%) but not all scenarios.* One may additionally add that previously computed scenarios, possibly using much coarser grids and less advanced models, would have resulted in similar or consistent projections. Such an approach might be criticized in some quarters as underselling the utility of scenarios, but it is our view that there is greater risk in communicating quantified uncertainty when the basis for that quantification is not clear. We should be clear that they are scenarios, and not forecasts. Regardless of how the information in an ensemble of scenarios is communicated, their core utility is that they provide planning tools based on the principles of physics and on possible, plausible and internally consistent ideas (in the sense of Schwartz 1991) about future.

Of course, assessments that are expressed using calibrated language using terms such as "confidence" and "likelihood", and that may be informed by different combinations of expert judgement and quantitative analysis, may also often be useful (IPCC 2007). Physical plausibility in the light of theoretical reasoning often adds robustness to the interpretation of climate change scenarios.

Acknowledgement Many thanks to Moritz Maneke of Norddeutsches Klimabüro, who supplied us with the regional scenarios.

References

Abramowitz G (2010) Model indepdence in multi-model ensemble prediction. Austral Meteorol Ocean J 59:3–6

Allen M, Stott P, Mitchell J, Schnur R, Delworth T (2000) Quantifying the uncertainty in forecasts of anthropogenic climate change. Nature 407:617–620

Bray D, von Storch H (2009) Prediction or projection? The nomenclature of climate science. Sci Commun 30:534–543. doi:10.1177/1075547009333698

Conover WJ (1971) Practical nonparametric statistics. Wiley, New York, p 462

IPCC (2007) Climate change 2007: the physical science basis. In: Solomon S et al (eds) Contribution of working group I to the 4th assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, p 996

Meinke I, Gerstner E, von Storch H, Marx A, Schipper H, Kottmeier C, Treffeisen R, Lemke P (2010) Regionaler Klimaatlas Deutschland der Helmholtz-Gemeinschaft informiert im Internet öber möglichen künftigen Klimawandel. Mitteilungen DMG, 02/2010, pp 5–7. http://www.norddeutscher-klimaatlas.de/klimaatlas/2071-2100/jahr/durchschnittliche-temperatur/norddeutschland.html (in German) (2010)

Murphy JM, Booth BBB, Collins M, Harris GR, Sexton DMH, Webb MJ (2007) A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. Philos Trans R Soc Lond A 365:1993–2028

Murphy JM et al (2009) UK climate projections science report: climate change projections. Met Office Hadley Centre, Exeter. http://ukclimateprojections.defra.gov.uk. Accessed 17 March 2011

Murphy JM, Sexton DMH, Barnett DN, Jones GS, Webb MJ, Collins M, Stainforth DA (2004) Quantification of modelling uncertainties in a large ensemble of climate change simulations. Nature 429:768–772

Nakicenovic N, Swart R (eds) (2000) Special report on emissions scenarios. Cambridge University Press for Intergovernmental Panel on Climate Change, Cambridge, UK

Rougier J, Goldstein JM, House L (2010) Assessing climate uncertainty using evaluations of several different climate simulators. Unpublished manuscript, p 32. http://www.maths.bris.ac.uk/~MAZJCR/mme2.pdf. Accessed 14 Dec 2011

Schnorbus MA, Bennett KE, Werner AT, Berland AJ (2011) Hydrologic impacts of climate change in the peace, Campbell and Columbia Watersheds, British Columbia, Canada. Pacific Climate Impacts Consortium, University of Victoria, Victoria, BC, p 157

Schwartz P (1991) The art of the long view. Wiley, New York, p 272

Tebaldi C, Knutti R (2007) The use of the multi-model ensemble in probabilistic climate projections. Philos Trans R Soc A 365:2053–2075. doi:10.1098/rsta.2007.2076

von Storch H, Zwiers FW (1988) Recurrence analysis of climate sensitivity experiments. J Climate 1:157–171

von Storch H, Zwiers FW (1999) Statistical analysis in climate research. Cambridge University Press, Cambridge, UK, p 484

Woth K (2005) Projections of North Sea storm surge extremes in a warmer climate: how important are the RCM driving GCM and the chosen scenario? Geophys Res Lett 32:L22708. doi:10.1029/2005GL023762