List of changes

We thank the reviewers once again for their helpful comments and suggestions. Hereafter we list the modifications made to the original manuscript. We also provided a "track-changes" version of the manuscript, where you can see all the modifications.

List of changes according the the comments by Reviewer 1

Major comment (A): The presentation of the economic framework is too short.

Additional references for the presentation of economic elements, specifically risk aversion and utility theory:
We added more references on those topics from pear-reviewed journals in the revised version of the manuscript: Krzysztofowicz (1986), Merz et al. (2009), Shorr (1966), Cerdá Tena and Quiroga Gómez (2008), Werner (2008), Fishburn (1989) and Pope and Just (1991). In particular, Fishburn (1989) provides a retrospective on vNM utility theory with many excerpts from the original book by von Neumann and Morgenstern (1944).

Using the above references, we extended the presentation of the economic framework (section 2 and also Appendix A) in order for it to be more informative for Earth scientists.

About the definition of risk aversion
We extended section 2 to explain risk aversion in greater details. Specifically, lines 25-29 on page 3 of now read: "Risk aversion" refers to an attribute of a decision maker who would be willing to pay a certain amount of money to remove any risk associated to a decision problem. The specific amount of money he or she is willing to pay for this is initially unknown and can be seen as an indirect measure of the magnitude of this aversion."

About the CARA utility function and reorganization of section 2
Section 2 was reorganised and Figure 1 was modified so that it is based (schematically) on the CARA utility function instead of on a generic function. We also expanded Appendix B, which describe the properties of the CARA utility function. First, we refer to Figure 1 in this Appendix, as you suggested. Also, the following was added (lines 3-5 page 21):
"The value of A reflects the decision maker's level of risk aversion. Specifically, the *Arrow-Pratt index of absolute risk aversion* is defined as

$$A(\mu) = \frac{-\mu''(\cdot)}{\mu'(\cdot)} \tag{1}$$

for all twice continuously differentiable function $\mu(\cdot)$. If $A(\mu) > A(\tilde{\mu})$, we say that the decision maker whose preferences are represented by μ is more risk-averse than a decision maker whose preferences are represented by $\tilde{\mu}$.
Using the parametric form: $\mu(x) = \frac{-1}{A}\exp\{-Ax\}$, we immediately see that $A(\mu) = A$. Since $A(\mu)$ is independent of x, we say that μ exhibits a constant absolute level of risk aversion."

This shows why an increase in A is equivalent to an increase in the level of risk-aversion, and why the level of risk-aversion is independent of the wealth.

How μ reflects the decision maker's preferences regarding uncertainty
Section 2 now explains in greater details the link between μ and the decision maker's preferences as well as the link between concavity and risk aversion. In particular, lines 5-11 page 5 now reads:
"To see why, consider the random variable $\tilde{c}$, and its expected value $\bar{c}$.[1] Since $\bar{c}$ is not risky, a risk-averse decision maker should prefer receiving $\bar{c}$ with certainty than receiving a random draw from $\tilde{c}$. That is: $U(\bar{c}) > U(\tilde{c})$, or $\mu(\bar{c}) > \sum_{m=1}^{M} p_m \mu(c_m)$, which is the definition of concavity. Note that we can also define $C > 0$, the amount of money that the decision maker would be willing to spend to remove the risk associated with $\tilde{c}$, as follows:

[1] Note that $\bar{c}$ can be thought as a degenerated random variable, taking the value $\bar{c}$ with probability 1.

$$\mu(\bar{c} - C) = \sum_{m=1}^{M} p_m \mu(c_m) \tag{2}$$

This argument extends directly to any change in risk: any risk-averse decision maker prefers less risky distributions, in the sense of mean-preserving second order stochastic dominance (Rothschild and Stiglitz, 1970). Figure 1 also presents a graphical version of the above discussion when there are only two states of nature."

Appendix A was also modified (page 20 line 23-27):
"To see more clearly the impact of risk-aversion on the optimal decision, suppose that μ is CARA, i.e. $\mu(x) = \frac{-1}{A}\exp\{-Ax\}$, and that $b = d$. Using the formula above and straightforward algebra, we find that an action is optimal if

$$p \geq \frac{\exp\{Ac\} - 1}{\exp\{Ad\} - 1} \equiv t(A) \tag{3}$$

as opposed to $p \geq c/d$ for the cost-loss ratio. One can verify that $t(A)$ is strictly decreasing with $\lim_{A\to 0} t(A) = c/d$. Then, this implies that, as risk aversion increases, the decision maker requires lower confidence level (for the realisation of the adverse event) in order to take an action. The limiting case, when the decision maker is risk neutral, gives the cost-loss ratio."

Major comment (B): How is the upper tail of the predictive distribution taken into account?

"States of the world"
The term "state of the world" is now precisely defined at page 4 lines 27 to page 5 line 2:
"The set of states of the world represent the set of realizations of $\tilde{c}$ for which the decision maker has preferences For instance, in Cerdá Tena and Quiroga Gómez (2008), there are only two possible states of the world: "adverse weather" and "non adverse weather". [2] In the case of flood forecasting systems, even if the streamflow values are continuous, the decision maker may only distinguish between a finite set of implied damages. This point is discussed further in section 4.2 where a finite number of "damage categories" are specified."

Missed events in the database and sufficiency of data to draw full conclusions:
The revised version of the manuscript includes a new figure (Figure 11) instead of previous boxplot of differences $Q_{fcst} - Q_{obs}$. It displays histograms reporting the number of events observed in each class of events, for each forecasting system. See also the text at page 17 lines 17-25. This figure shows, among other things, that all forecasting systems generally overforecast. Missed events are not a big issue on the Montmorency watershed. However, we do agree with you that this is an important issue for flood forecasting in general.

We also discuss the issues related to the length of the data base in greater details. See page 13 lines 21-24:
"On the one hand, it is expected that a longer record will provide a better empirical estimate of the true streamflow distribution. On the other hand, there can also be various sources of non-stationarity affecting the observed streamflow values over time (e.g. changing the measurement apparatus, climate change, land-use change, etc). Hence, even with a very long historical record, the true distribution of streamflow cannot be known with certainty. (Note that this also affects measures of quality, such as the CRPS.)"

Major comment (C): The discussion consider non scientific issues which some hydrologists and forecasters can disagree with.

The discussion was revised according to the comments from both reviewers. First, a portion of text was moved from section 4.3 to the Discussion (line 27 page 18 to line 7 page 28, new version) according to a suggestion from Reviewer 2.
We will not recopy the new discussion here as it is long, but you can see the changes in the "track-changes" version of the manuscript. Those changes follow your suggestions, namely:

[2] **vNM utility functions can also account for an infinite number of states of the world. In such case, one would have: $U(\tilde{c}) = \int \mu(c) f(c) dc$, where f is the pdf of $\tilde{c}$.**

- We removed any part of text that could have suggested that forecasters should modify their true belief.
- We emphasized that "In any case, it is capital to recall that the role of the forecaster is to issue the best possible streamflow forecast given their knowledge of the situation and available model and data." (page 18 lines 21-24)
- We moderated the conclusion, to avoid going to far given the case study. To do this, and limited to your comments, as expressed on page C4 of your review.
- We emphasize that in our study, we consider a well-trained decision maker, but we mention the issue of training and also potential cognitive biases. We also added references (in hydrology) in which the issue of training is discussed (Ramos et al., 2013; Demeritt et al., 2010; Doswell, 2004).

Lastly, we added more precision regarding the fact that forecasts users (and people in general) are not aware of their precise level of risk aversion. Most people have a general idea of their behaviour toward risky situations (i.e. for instance a person who likes to gamble can assume that she likes taking risks, at least to some extent), but it is hard to pinpoint the precise value of A. For this reason, lines 7-9 page 18 now reads:

"The "real" level of risk aversion for the decision maker for flood emergency measures along the Montmorency River remains unknown due to the insufficient record of decisions and associated spending. However, it can be reasonably assumed that they are highly risk-averse (Claude Pigeon, personal communications)."

To illustrate this further, we can mention that, on several occasions during phone calls related to this work, Mr. Pigeon worried about the potential occurrence of deaths related to flooding. We asked if there has ever been deaths caused directly or indirectly by floods on the Montmorency River. His answer was: "No, but there *could* be, someday". Of course this is a very specific example, and maybe extreme, but it is similar to the example of the Minister from Prague in the manuscript. It is probably safe to say that most decision makers involved in flood mitigation will "play on the safe side" and "not take any chance" with peoples' property (and life!). While this doesn't provide a precise value for A in a utility function, this is indicates a risk-averse behaviour, which cannot be represented by the cost-loss ratio.

Appendix A was also further modified to include a simple mathematical demonstration of the impossibility of considering risk-aversion in the cost-loss ratio framework (page 20 line 23-27, already mentionned above). A similar demonstration, is also performed in Cerdá Tena and Quiroga Gómez (2008).

Detailed comments

We will simply refer to the line of the modified manuscript addressing each comment

Your comments from Page 2:

Line 4: "uncertainty *assessment of* hydrological forecasts conveys important information for decision makers' rather 'uncertainty in hydrological forecasts conveys important information ..."
Page 2 line 5

Line 7: I agree that the analog forecasts are of common use. However, why quoting this approach first? Is it used by the DEH? How is it relevant for this article? I suggest the authors list the most importance uncertainty sources (in a sorted way) and then present the methodologies which can be used to deal with them. The link between analog forecasts and then ensemble forecasts (line 13) is not clear.
Page 2 in general, and specifically lines 7-14 for sources of uncertainty.

Line 13 ("ensemble forecasts are superior to deterministic ones"): do the authors focus on ensemble forecasts or is it true as well for probabilistic forecasts? (ensemble forecasts being used as "proto" or substitute of probabilistic forecasts, since probabilistic information is drawn from this kind of forecasts).
Page2 Line 20

Lines 16-18: I agree that economic value assessment is not straightforward. However, assessing a forecast system by comparing forecasts with corresponding observations is not straightforward either.

Indeed, there is not one quality but different qualities (especially for probabilistic (and then ensemble) forecasts). Different end-users would give different weights to these qualities (since they have specific applications).
Page 2 Lines 28-29

Line 26 ("which does not fully exploit the information about forecast uncertainty"): what does "fully" mean here? Verkade and Werner (2011) do take explicitly into account the uncertainty.
Page 3, lines 1-2.

Line 31: check spelling (Neumann / Newman)
Page 3, line 6

Line 32: since the proposed framework is based on the von Neumann and Morgenstern utility function, more references are needed than a single one of 1944. Another reference is given further (page 3, line 30). But it is a book, which may be a "classic" in the economic community, but not the easiest reference to find and read by a hydrologist.
Indeed. Please see our answer to the general comment A.

Your comments from Page 3:

Lines 5 and 6: this sentence provides some conclusions of the article. Why here in the introduction?
Removed

Line 12 ("Results are presented and discussed in section 6"): in sections 6 and 7. Line 18 ("Most importantly"): do the authors mean "More importantly"?
Page 3, line 24

Line 23: check English (spelling for "weighting")
Page 4 line 8

Line 30: see comment for page 2, line 32.
Additional references were added.

Your comments from Page 4:

Line 5 ("the curvature of the function mu reflects the decision maker's preference regarding uncertainty"): why? Some references would be gratefully welcome.
Indeed. Please see our answer to the general comment A.

Line 9: isn't a reference to Fig. 1 missing here?

Figure 1 modified. It is now based on the CARA utility function.

Line 20: check English ("teh")
Page 6 line 7

Line 21: check the numerotation of tables (table 3 is referred to before table 1 and 2)
Done!

Your comments from Page 5:

Lines 32-33 (and lines 1-2 page 6): I did not understand why the HYDROTEL file system is useful for the reader. Are these technical details significant for this study or may they be avoided?
Removed (page 7)

Your comments from Page 6:

Lines 31-33: I am not sure that I understood correctly. Is the meteorological forecast ensemble used here computed by the meteorological service of Canada but taken from the TIGGE dataset (for some practical reasons)? If so, it might be clearer if stated this way.
Page 8 lines 8-9.

Your comments from Page 7:

Lines 10 & 11 ("Thiboult et al. (2016) showed that the [...]"): please be more specific (for this catchment? For this area?...)
Page 8, line 19.

Line 16: the additive coefficients for temperature inputs and the multiplicative coefficients for precipitation inputs are huge and I assume that they are much larger than the uncertainty for these inputs. Is the whole range used in practice? Is this manual 'tuning' used for more than reducing the input uncertainty in getting a best guess? Some discussion would be useful here.
Yes they are huge. They are the true operational limits at the DEH. However, it is worth emphasizing that the goal of those perturbations on precipitations and temperature is to (indirectly) affect state variables (soil moisture, snow water equivalent) and correct model uncertainties. They are not intended as to reflect the true uncertainty on precipitations and temperature. The goal of this manual tuning is indeed to obtain a best guess regarding the initial state of the watershed (under the assumption that the state variables of the model accurately reflect the state of the watershed). However, it might reassure the reviewer (as well as everybody else) to know that those huge limits for perturbations are rarely reached. In our study, the multiplicative coefficient applied to precipitation varied between 0.5 and 2.5. Most additive coefficients for temperature varied between -3 and +2.5, with occasional large coefficient (up to -7 and +7 on 2-3 occasions). Precisions regarding what perturbations were really applied and the limits that were permitted were added in the revised version of the manuscript.

Page 8 lines 27 to Page 9 line 1.
The EnKF that is implemented here follows Thiboult et al. (2016) and Mandel (2006). M is the model error covariance matrix, computed before data assimilation, at each time step of the sequential data assimilation. As such, it is not updated, as it is the model's state variables that are updated according to equation (3). Of course, updating the state variables will affect the model outputs, hence M at the following time step. Thus, in our specific implementation state variables are indeed updated, but not from an open loop simulation. The base line simulation here is the manually assimilated run. This base line simulation is good, but cruelly lacks dispersion. In that context, the purpose of the EnKF is only to consider uncertainty associated to state variables and not to improve the first guess estimate of state variables. The parameters of the EnKF were not fine tuned as in many studies (such as Thiboult and Anctil, 2015), for instance. Random perturbations added to temperature were drawn from uniform distributions U[-8,+8]° and U[0.5, 1.5] (multiplicative) for precip. This choice is coherent with the way the manual data assimilation was performed, but could certainly be improved. For instance, normal error distributions are most commonly used and the spread of those distribution is calibrated until good agreement with observation is achieved. Again, the goal of the EnKF here is to add spread around best estimate of state variables, in a controlled and systematic manner. We consider that further refinement of the EnKF is outside the scope of our study.

Your comments from Page 8:
Line 16: does 's' include the cost of the forecasting system (independently from the money spent for risk mitigation)?
Page 9, lines 15-19.

Line 21: may the author provide some figures (orders of magnitude?) or some plots?
Unfortunately we confirmed with the Civil Security of Ville de Québec that we can't, as these are confidential.

Line 28 ("these represent relatively small levels of risks of aversion"): may the authors provide some references?
Page 10 line 26.

Line 28 ("it is shown that they lead to qualitative changes in the decision makers"): here again, some

references would help the non specialist reader.
Page 10 line 26-28

Your comments from Page 9:

Line 25: as a non specialist, I was amazed by the range of the psi factor (1.5 to 10). Is this usual?
The range of ψ captures two important aspects. First, in 2014, the civil security spent around 3.5 times more than the realized material damages. This reflects the fact that (perhaps obviously) the decision maker also consider immaterial damages. Since it is extremely hard to evaluate immaterial damages, we let ψ vary to (very) large values. We actually also performed simulations for values much higher than 10. They are not displayed in the current version of the manuscript as the analysis would remain the same, but the graphs would be harder to read.
We believe that $\psi = 10$ is a reasonable value. Recall that immaterial damages include any damage that cannot be easily expressed in monetary value. Those include losses in the "quality of life", avoiding law suits (including the associated bad press)... and can therefore be quite high.

Your comments from Page 10:

Line 16: the accuracy of forecasts is inversely related to lead time. Is it inversely proportional to it?
Page 11, line 31.

Line 29: I am not sure that I understood the division of parameter β_m. Why all factors $(2, 1.75, 1.5, ...)$ are larger than 1? I would have expected weights whose sum is 1.
Those factors reflect the benefit of early warning. They are not weights. The baseline is the 1-day ahead warning, so any early warning should be more beneficial. In practice, this reflects the fact that the population has time adjust (pack, empty their basements, arrange visits to their relatives...) before being evacuated.

Your comments from Page 11:

Line 17: why 'then'? First results provided are the hydrographs (Fig. 3) on which doing the visual inspection.
Page 13, lines 4-7.

Your comments from Page 12:
Lines 24-25: I suggest that the information of Appendix C comes in the main text (it is necessary for the reader).
Page 12 line 29 to page 13 line 1.

Your comments from Page 13:
Lines 7 & 8 ("This figure shows that for 1-day forecasts, those based on meteorological ensembles and dressed deterministic forecasts have similar spread"): this is not obvious for me.
Page 14 lines 27 to page 15 line 1.

Line 16 ("For very short lead times, the dressed deterministic forecasts outperform those based on meteorological ensembles"): some discussion (interpretation) would be appreciated on this (common) behaviour.
Page 15, lines 9-10.

Line 20: in practice, how does the DEH deal with the very "jumpy" ensemble curves? Are they used by operational forecasters?
No. The DEH doesn't use forecasts based on meteorological ensembles. They use the dressed deterministic forecasts, which are not so "jumpy". This is mentioned on page 7 (line 27-29).

Your comments from Page 14:
Lines 16 & 17 ("for higher level of risk aversion [...], the decision maker SHOULD prefer the 'no forecast' situation for low levels of Psi"): doesn't the modal verb convey a notion of duty? (you are right if you do what you should do). I would rather write that the forecasting system has no (economic value) or usefulness for highly risk-averse users.
Page 16, lines 10-11

Your comments from Page 16:
Line 12 ("The economic value of a forecasting system is necessarily dependent on the level of risk aversion of the decision maker"): first, it is more the economic value of the forecasts (you have to deduced its cost to get the value of the forecasting system). Then, even if I agree on the fact that it is very common (if not always), is this "necessary"? Can it be shown?
This sentence was removed during the rewriting of the Discussion section (page 18).

Lines 23-26: this paragraph has to be emphasized. Morevore, communicating the forecasts in a way that the end-users would perfectly understand is a key, but it is totally different from 'overforecast'.
Please see our answer to major comment C and also the revised Discussion (page 18)

Your comments from Page 18

Appendix A: it is referred before section 2.1 but it uses the concepts presented in this section.
Page 4, lines 7-8.

Appendix B. Where is Fig. 1 called?
Figure 1 was modified to avoid confusion and now represents the CARA utility function. Text has been modified accordingly. See also our answer to major comment A.

Your comments from Pages 23 & 24:
Tables 1 and 2 might be merged since their comparison is highly teachingful.
We agree and this was done. Table 1 now includes data from both those previous tables.

Your comments from Page 25
Table 3 could usefully be replaced by a plot of monthly values (if data is available)
We agree. Table 3 was removed. It is now replaced by Figure 2.

Your comments from Page 27:
Fig. 1: why is not the utility function plotted for negative values? Because if c $<$ 0, then there is no 'interest' then the utility is 0? If so, why to use it with negative values in appendix A (for example, mu(-d))?
Figure 1 now includes negative values of c. Indeed, one of the advantages of working with CARA utility functions is that they are defined for any value of c, and not just for positive values.

List of changes according the the comments by Reviewer 2

Again, we will simply refer to the line of the modified manuscript addressing each comment unless it is necessary to do otherwise. More detailed replies were provided in our initial response to your comments.

Abstract

- No abbreviations should be used in the abstract without explanations.
CARA: Page 1, line 9; HEPEX: Page 2, line 2; SWE: removed; BV3C: Page 7, line 4; THORPEX: Footnote page 8.

- A sentence summarizing the main conclusions of the work should be added.
Page 1 line 23-24: "Hence, post-processing forecasts to avoid over-forecasting could help improving both the quality and the value of forecasts."

Introduction
Again, abbreviations are not well explained. Please provide the full term when the abbreviation is used for the first time. Please check the whole paper.
Done

p. 2, line 7/8: What does this mean? Could you provide examples?
According to a comment by Reviewer 1, all references to analog forecasting systems have been removed from the revised version of the manuscript.

In general, a more structured review of the literature on uncertainty is missing. For example, different types of uncertainty (epistemic versus aleatoty/natural uncertainty) could be distinguished since they may have different effects on decisions and decision makers because epistemic uncertainty can be reduced by better data or models while aleatory uncertainty cannot. Later in the paper, this should also be discussed in the context of the study.
More references regarding the types of uncertainty were added and discusses on page 2, lines 7-14.

The von Neumann and Morgenstern utility function should already be briefly explained in the introduction (p. 2, line 31/32)
Please see our answer to comment A from Reviewer 1.

p. 3, line 2: delete "forecast" once.
Done!

Section 2

If you use a section 2.1 there should also be a section 2.2. One subheading does not make sence. Consider to delete the headline.
The heading of previous subsection 2.1 was deleted (page 3).

The economic model and the utility functions should be better explained. The content of the chapter referenced in line 30 (p. 3) should be briefly summarized.
Please see our answer to major comment A by Reviewer 1. He/she also had many questions and comments regarding those topics and asked for additional references. We indicated pages and lines corresponding to modifications in our answer.

A paragraph that bridges this section to the next should be added
Page 5 lines 16-17.

Starting on p. 4: Check the numbering of the equations; add numbers to all equations on p.4,9 and 12.
Thank you for pointing this out. All equations are now numbered

Section 3

Typo in line 20 (p. 4)
This has been corrected.

p. 4, line 28/29: consider rephrasing, check logic
Those lines currently read: "The response time of the watershed is rapid (12 hours). The return period of damaging floods is also short. This makes emergency evacuation and flood damage a common occurrence for riverside". We would like more precisions on what to clarify. First sentence means that floods appear rapidly. Second sentence means that floods happen often. This is why it is important to have flood forecasts and an emergency plan for this particular watershed.

In Table 1, the potential damage should be added fpr each return period.
The revised version of the manuscript includes histograms reporting the number of events observed in each class of events, for each forecasting system. We believe that this information will be more in line with the general framework of the paper. In addition, we are worried that displaying the values derived from the flow-damage curve provided in Leclerc et al. (2000), which would be gross approximations, could lead the reader to put too much confidence in those estimates.Note also that Leclerc et al. (2000)'s report (though in French) is freely accessible on the Internet.

p. 5, line 8/9 consider rephrasing ("cause" is used twice in this short sentence)
Page 6, lines 17-18.

This is unclear. The calibration performed by the DEH should be explained (as well as the meaning of DEH - see my comment on the use of abbreviations)
Page 7, lines 7-10.

Again, there shouldn't be a section 3.3.1 only. Please reorganize the text.
Contrarily to our initial answer, we decided instead to add a section 3.3.2. Please see page 8.

Section 4

p. 8, line 14: The use of 12 categories should be justified or better explained.
Please see page 10, lines 2-6

p. 8, line 19-21: The content and use of the data for the 2014 flood is unclear. Please add some information.
Please see page 11, lines 10-13 and also page 12 lines 12-15.Unfortunately, if we can share the value of the streamflow (825 m3/s), our confidentiality agreement with the civil security prevent us to communicate the amount spent.

p.9, line 4-6: The basis/source of the mentioned losses is unclear. Please explain how these values were derived. In line 27, a damage curve of Leclerc et al. (2001) is mentioned. This comes too late and too vague. Explain how the curve looks like, whether it is applicable in the catchment under study or/and whether and how is was adpated to your case study.
Greater details were added on page 10 line 32 to page 11 line 3.

p.9, line 4 and line 10: consider using "losses" instead of "damages"
We choose to use "damage" instead of "loss" in order to distinguish from the usual use of the term "loss", as in "cost-loss ratio". As described in Appendix A, under risk aversion the two are not necessarily equivalent. We prefer using "damage" representing the actual, incurred, damages.

p. 10, line 3 to 15: Most of this should be shifted to the dicsussion section.
We agree, this part was moved to the discussion. Please see the bottom of page 18.

In general, the section 4.3 is somewhat unclear and contains too many issues for discussion. Consider to shorten it to the main point that are necessary for the model application. This is related to your previous comment. We agree that some items should be moved to the discussion. It will be done in the revised version of the manuscript. **Section 5**

p. 12, line 3: discuss how the true distribution of streamflow could be determined or whether it is possible to check the validity of the used distribution.

It is actually not possible to determine the true distribution of streamflow. One can only approach it by using the available historical record. We elaborate on this issue on page 13, line 19-24.

Section 6

p.13, line 22: What do you mean by "sharpness"? Accuracy?
Please see page 15, lines 16-19.
What do you refer to when you mention "relatively rare and comparatively small flood events"?
You are right that our statement was imprecise. We will add more details in the revised version of the manuscript.We mean that the "usual" flood events for the Montmorency River are much less dramatic than the predicted ones (looking at the upper tail of the predictive distribution).We added a precision at page 16, lines 17-20.

Section 8

p.17, line 19: typo "AND in terms..."
Corrected!

Figure 3, 4 and 10: Explain the abbreviations in the figure caption
We would really prefer to leave the abbreviation in the figures for two reasons (1) The acronyms (EnKF and CRPS) are already defined in the text, before the figures (please see our answer to your comments about the introduction) and (2) The use of acronyms in figure titles allows for those titles to remain relatively short, which in our opinion is better for ease of reading.

References

Cerdá Tena, E. and Quiroga Gómez, S.: Cost-Loss Decision Models with Risk Aversion, 01, Instituto Complutense de Estudios Internacionales, 2008.

Demeritt, D., Nobert, S., Cloke, H., and Pappenberger, F.: Challenges in communicating and using ensembles in operational flood forecasting, Meteorological Applications, 17, 209–222, 2010.

Doswell, C.: Weather forecasting by Humans - Heuristics and Decision Making, Weather and Forecasting, 19, 1115–1126, 2004.

Fishburn, P.: Retrospective on the Utility Theory of von Neumann and Morgenstern, Journal of Risk and Uncertainty, 2, 127–158, 1989.

Krzysztofowicz, R.: Expected utility, benefit, and loss criteria for seasonal water supply planning, Water Resources Research, 22, 303–312, 1986.

Leclerc, M., Heniche, M., Secretan, Y., and Ouarda, T.: Travaux d'atténuation des risques de crue à l'eau libre de la rivière Montmorency dans le secteur des îlets – PHASE 2. Mise à jour de l'analyse hydrologique, idmensionnement des travuax d'atténuation et analyse de l'impact sur les risques résiduels de dommage aux résidences., Tech. Rep. R555, INRS-Eau, Quebec, 2000.

Mandel, J.: Efficient implementation of the Ensemble Kalman Filter, Tech. Rep. R1416, University of Colorado at Denver and Health Sciences Center, Denver, 2006.

Merz, B., Elmer, F., and Thieken, A.: Significance of "high probability/low damage" versus "low probability/high damage" flood events, Natural Hazards and Earth System Sciences, 9, 1033–1046, 2009.

Pope, R. and Just, R.: On testing the structure of risk preferences in agricultural supply analysis, Agricultural Journal of Agricultural Economics, 73, 743–748, 1991.

Ramos, M.-H., van Andel, S., and Pappenberger, F.: Do probabilistic forecasts lead to better decisions?, Hydrology and Earth System Sciences, 17, 2219–2232, 2013.

Rothschild, M. and Stiglitz, J. E.: Increasing risk: I. A definition, Journal of Economic theory, 2, 225–243, 1970.

Shorr, B.: The cost/loss utility ratio, Journal of Applied Meteorology, 5, 801–803, 1966.

Thiboult, A. and Anctil, F.: On the difficulty to optimally implement the Ensemble Kalman filter: An experiment based on many hydrological models and catchments, Journal of Hydrology, 529, 1147–1160, 2015.

Thiboult, A., Anctil, F., and Boucher, M.-A.: Accounting for three sources of uncertainty in ensemble hydrological forecasting, Hydrology and Earth System Science, 20, doi:10.5194/hess-20-1809-2016, 2016.

von Neumann, J. and Morgenstern, O.: Theory of games and economic behavior, vol. 60, Princeton University Press Princeton, 1944.

Werner, J.: risk aversion, in: The New Palgrave Dictionary of Economics, edited by Durlauf, S. N. and Blume, L. E., Palgrave Macmillan, Basingstoke, 2008.

Moving beyond the cost-loss ratio: Economic assessment of streamflow forecasts for a risk-averse decision maker

Simon Matte[1], Marie-Amélie Boucher[1], Vincent Boucher[2], and Thomas-Charles Fortier Filion[3]

[1]Dept. of Applied Sciences, Université du Québec à Chicoutimi, 555, boulevard de l'Université, Chicoutimi, G7H 2B1, Canada

[2]Dept. of Economics, Université Laval, 1025, avenue des Sciences-Humaines, Québec, G1V 0A6, Canada

[3]Québec Government Direction of Hydrologic Expertise, 675, boul. René Lévesque Est., Québec, G1R 5V7, Canada

Correspondence to: Marie-Amélie Boucher (marie-amelie_boucher@uqac.ca)

Abstract.

A large effort has been made over the past 10 years to promote the operational use of probabilistic or ensemble streamflow forecasts. Numerous studies have shown that ensemble forecasts are of higher quality than deterministic ones. Many studies also conclude that decisions based on ensemble rather than deterministic forecasts lead to better decisions in the context of flood mitigation. Hence, it is believed that ensemble forecasts possess a greater economic and social value for both decision makers and the general population. However, the vast majority, if not all, of existing hydro-economic studies rely on a cost-loss ratio framework that assumes a risk-neutral decision maker. To overcome this important flaw, this study borrows from economics and evaluates the economic value of early warning flood systems using the well-known **Constant Absolute Risk Aversion (CARA)** utility function, which explicitly accounts for the level of risk aversion of the decision maker. This new framework allows for the full exploitation of the information related to a forecasts' uncertainty, making it especially suited for the economic assessment of ensemble or probabilistic forecasts. Rather than comparing deterministic and ensemble forecasts, this study focuses ~~rather~~ on comparing different types of ensemble forecasts. There are multiple ways of assessing and representing forecast uncertainty. Consequently, there exists many different means of building an ensemble forecasting system for future streamflow. One such possibility is to dress deterministic forecasts using the statistics of past error forecasts. Such dressing methods are popular among operational agencies because of their simplicity and intuitiveness. Another approach is the use of ensemble meteorological forecasts for precipitation and temperature, which are then provided as inputs to one or many hydrological model(s). In this study, three concurrent ensemble streamflow forecasting systems are compared: simple statistically dressed deterministic forecasts, forecasts based on meteorological ensembles and a variant of the latter that also includes an estimation of **state** variable uncertainty. This comparison takes place for the Montmorency River, a small flood-prone watershed in south central Quebec, Canada. The assessment of forecasts is performed for lead times of one to five days, both in terms of forecasts' quality (relative to the corresponding record of observations) and in terms of economic value, using the new proposed framework based on the CARA utility function. It is found that the economic value of a forecast for a risk-averse decision maker is closely linked to the forecast reliability in predicting the upper tail of the streamflow distribution. **Hence, post-processing forecasts to avoid over-forecasting could help improving both the quality and the value of forecasts.**

1 Introduction

More than fifteen years after its advocation by Krzysztofowicz (2001) and more than a decade after the creation of the **Hydrologic Ensemble Prediction EXperiment (HEPEX)** community (Franz and Ajami, 2005; Schaake et al., 2007), *the case for probabilistic forecasting in hydrology* has been accepted by many researchers and practitioners across the world: uncertainty ~~in~~ **assessment of** hydrological forecasts conveys important information for decision makers and therefore should be quantified and be considered as part of the forecast.

~~However, as there exists multiple sources of uncertainty in hydrological processes, there also exists many means of assessing this uncertainty and building an ensemble to represent this uncertainty.~~ **Beven (2016) distinguishes aleatory uncertainty, that originates from data only and possess stationary statistical characteristics, from various types of epistemic uncertainties. Epistemic uncertainties can arise from a lack of knowledge regarding the system's dynamics, from a lack of knowledge regarding the relevant forcings for the modeling process and also from disinformation in the data. More broadly speaking, as discussed in Juston et al. (2013), uncertainty in hydrological forecasting mainly originates from data and models (atmospheric and hydrologic). The most important sources of uncertainty in short-term hydrological forecasting are structural uncertainty (choice of a particular hydrological model structure), state variable uncertainty and parameter uncertainty, which are both linked to the availability and quality of hydro-meteorological data, and meteorological forecasts uncertainty. The latter gain in importance gradually as the forecasting horizon increases**. ~~It is common for operational agencies to resort to analog forecasts (e.g. Hamill and Whitaker, 2006; Diomede et al., 2008; Marty et al., 2012). There are many variants of analog forecasting systems, but all rely on an assessment of past forecasting errors to build the ensemble.~~

However, as there exists multiple sources of uncertainty in hydrological processes, there also exists many means of assessing this uncertainty and building an ensemble to represent this uncertainty. It is possible, for instance, to produce streamflow ensemble forecasts from meteorological ensemble forecasts used as inputs to at least one previously calibrated hydrological model. Deterministic forecasts can also be "dressed" using past error statistics. ~~It is also possible to produce streamflow ensemble forecasts from meteorological ensemble forecasts used as inputs to at least one previously calibrated hydrological model. Additional sources of uncertainty can be accounted for, in particular state variable uncertainty or parameter uncertainty.~~

While there is a general agreement among the global scientific community that ensemble **and probabilistic** forecasts are superior to deterministic ones (e.g. Jaun et al., 2008; Velazquez et al., 2010; He et al., 2013, and many others), there remains no consensus regarding the best means of obtaining an ensemble of streamflow forecasts (i.e. constructing the ensemble). There has **also** been **an** increased interest over the last few years in regards to assessing the economic *value* of forecasts. ~~Although t~~ **T**he quality of a forecasting system can be assessed by comparing forecasts for different lead times with corresponding observations**. Forecasts quality can be further decomposed into different attributes (e.g. resolution, sharpness, discrimination...) that can be weighted differently depending on specific applications.** ~~, f~~ **F**orecasts *value* **also** depend on the specific applications ~~and its assessment is not always straightforward (Katz and Murphy, 1997).~~ In particular, the usefulness of a forecast is

inherently linked to the decision maker's ability to adapt their behaviour to the information provided. **Neither the assessment of forecasts *quality* and *value* are straightforward and sometimes the relationship between the two is not obvious either.**

In the case of hydropower production, forecast values can be assessed using sophisticated decision-making models based on stochastic dynamic programming in an operational research framework (e.g. Boucher et al., 2012; Carpentier et al., 2013; Côte and Leconte, 2016). Early flood warning is another very important application for streamflow forecasts and a decision problem entirely different from the optimization of hydropower production. Hydrologists most often, if not always, assess the value of streamflow forecasts for early flood warning using the cost-loss framework (e.g. Murphy, 1977; Richardson, 2000; Roulin, 2007; Verkade and Werner, 2011), which ~~does not fully exploit the information about forecast uncertainty conveyed by the predictive distribution. In particular, it~~ does not account for the decision maker's *risk aversion*, i.e. the fact that, given the opportunity, a decision maker would be willing to spend money (or resources) to reduce the amount of uncertainty they face. This is discussed formally in section 2 below.

This study considers the evaluation of the economic value of early warning flood systems, from the point of view of the decision maker, with explicit consideration of risk aversion. This alternative framework is based on the use of the von ~~New-man~~**Neumann** and Morgenstern (vNM) utility function (von Neumann and Morgenstern, 1944), which is widely used in economics but rarely in hydrology.[1] To the best of our knowledge, our study represents the first attempt at accounting for risk aversion in the assessment of the economic value of streamflow forecasts for early flood warning. This new framework is used to assess the economic value of three concurrent streamflow ensemble forecasting systems in a case study for the Montmorency River, a flood-prone watershed in south central Quebec, Canada. Five day statistically dressed deterministic forecasts ~~forecasts~~ for this watershed have been issued operationally since 2008 by the Direction de l'Expertise Hydrique (DEH), a Quebec provincial agency. These forecasts are used for early flood warning and emergency response by the civil security bureau of Quebec City. ~~It is found that for risk-averse decision makers, the dressed deterministic forecasts have a higher economic value since they provide the most accurate prediction of the upper tail of the distribution.~~

In section 2, some concerns regarding the cost-loss ratio are raised and an alternative framework is presented. Section 3 describes the context of the case study, namely the specifics of the Montmorency River watershed, the current flood forecasting system based on dressed deterministic forecasts as well as the early flood warning mechanism in place. Two variants of a concurrent flood forecasting system are detailed in section 3.3. The economic model is presented in section 4. Performance assessment metrics, both in terms of forecasts quality compared to observations, and in terms of economic value, are presented in section 5. Results are presented ~~and discussed~~ in section 6 **and discussed in 7**. Conclusions are drawn in section 8 along with suggestions for future improvement of the proposed economic model.

[1] **Exceptions include Krzysztofowicz (1986) for seasonal water supply planning and Merz et al. (2009) for flood events, although Merz et al. (2009) use risk indicators and not vNM utility functions.**

2 The economic model and the limits of the cost-loss ratio

The cost-loss ratio decision model (Murphy, 1977; Katz and Murphy, 1997; Richardson, 2000) is a simplified framework used in numerous hydro-meteorological studies to assess the economic value of forecasts (Roulin, 2007; Abaza et al., 2014; Verkade and Werner, 2011, among many others). As pointed out by Zhu et al. (2002), this approach is only the simplest one out of a much larger range of options. ~~Most~~**More** importantly, a classical cost-loss ratio decision model disregards the role of risk aversion **(e.g. Shorr, 1966; Cerdá Tena and Quiroga Gómez, 2008). "Risk aversion" refers to an attribute of a decision maker who would be willing to pay a certain amount of money to remove any risk associated to a decision problem. The specific amount of money he or she is willing to pay for this is initially unknown and can be seen as an indirect measure of the magnitude of this aversion.**

As discussed by Cerdá Tena and Quiroga Gómez (2008), risk-aversion is very common, and most decision makers are risk-averse when the stakes are high. In their paper, they illustrate how disregarding risk aversion can sometimes lead to misleading conclusions regarding the value of information (such as meteorological or hydrological forecasts). Their framework also involves the Constant Absolute Risk Aversion utility function (see section 2). However, the context of their application and the rest of their economic model is different from ours.

In a simple cost-loss ratio, the decision model follows a contingency table that allows for binary decisions, with known associated costs. When applied to ensemble forecasts, decision-making according to the cost-loss ratio framework is based solely on a probability threshold associated to the material consequences of the event of interest (e.g. a flood event), regardless of the ensemble spread (uncertainty). Appendix A illustrates a technical presentation **that builds on the concepts presented in this section**. Including the concept of risk aversion in the decision model is not only more realistic, but allows ~~weighing~~**weighting** the ensemble members differently depending on the level of risk aversion. For instance, a risk-averse decision maker will give more importance to the forecasts members in the upper tail of the predictive distribution (i.e. highest streamflow values).

~~The next section describes the formal decision-making process.~~

In economics, "utility" is an ordinal notion that reflects the decision maker's preferences over a set of possible outcomes. **Preferred outcomes lead to greater utility values.** In the context of random outcomes, the most popular class of utility functions is the von Neumann and Morgenstern (vNM) utility function, as introduced in von Neumann and Morgenstern (1944).

Fishburn (1989) provides a retrospective on von Neumann and Morgenstern theory. He enlightens the remarkable impact this theory had on the subsequent development of economic theories and also clarifies some of its limits. There exists a immense amount of literature regarding the application of vNM utility theory in many different fields. For instance, Pope and Just (1991) compare different types of utility functions to represent preferences of farmers for potato acreage. Although we could not find previous work in hydrology where risk-aversion is considered in the assessment of the economic value of forecasts, Krzysztofowicz (1986) and Merz et al. (2009) acknowledge its importance. Shorr (1966) attempts a reconciliation of the cost-loss ratio framework with utility theory in the simple context of crop protection.

The interested reader is referred to Chapter 6 in Mas-Colell et al. (1995) for more details as well as the axiomatic foundations of vNM utility functions.[2]

The vNM utility function of a decision maker regarding a real-valued random outcome $\tilde{c}$ (e.g. money) is given by:

$$U(\tilde{c}) = \sum_{m=1}^{M} p_m \mu(c_m) \tag{1}$$

where $m = 1, ..., M$ are the different "states of the world", p_m is the probability of state m, and c_m is the realization of the random outcome $\tilde{c}$ in state m. The function $\mu(\cdot)$ is assumed to be non-decreasing.

The set of states of the world represent the set of realizations of $\tilde{c}$ for which the decision maker has preferences For instance, in Cerdá Tena and Quiroga Gómez (2008), there are only two possible states of the world: "adverse weather" and "non adverse weather". [3] **In the case of flood forecasting systems, even if the streamflow values are continuous, the decision maker may only distinguish between a finite set of implied damages. This point is discussed further in section 4.2 where a finite number of "damage categories" are specified.**

The curvature of the function $\mu(\cdot)$ reflects the decision maker's preference regarding uncertainty. If $\mu(\cdot)$ is concave, the decision maker is risk-averse; if it is linear, the decision maker is risk-neutral; if it is convex, the decision maker is risk-seeking. ~~Figure displays typical utility curves for risk-averse, risk-neutral and risk-seeking individuals. This can be seen in Figure~~ 1. **To see why, consider the random variable $\tilde{c}$, and its expected value $\bar{c}$.**[4] **Since $\bar{c}$ is not risky, a risk-averse decision maker should prefer receiving $\bar{c}$ with certainty than receiving a random draw from $\tilde{c}$. That is: $U(\bar{c}) > U(\tilde{c})$, or $\mu(\bar{c}) > \sum_{m=1}^{M} p_m \mu(c_m)$, which is the definition of concavity. Note that we can also define $C > 0$, the amount of money that the decision maker would be willing to spend to remove the risk associated with $\tilde{c}$, as follows:**

$$\mu(\bar{c} - C) = \sum_{m=1}^{M} p_m \mu(c_m) \tag{2}$$

This argument extends directly to any change in risk: any risk-averse decision maker prefers less risky distributions, in the sense of mean-preserving second order stochastic dominance (Rothschild and Stiglitz, 1970). Figure 1 also presents a graphical version of the above discussion when there are only two states of nature.

This study focuses on a well-known parametric family for $\mu(\cdot)$ known as the Constant Absolute Risk Aversion (CARA) function, given by Eq. 3.

$$\mu(c) = \frac{-\exp(-Ac)}{A} \tag{3}$$

where A is the risk aversion of the decision maker. A is strictly positive for risk-averse individuals and strictly negative for risk-seeking individuals. For positive value. the level of risk aversion increases when A increases.

[2] See also Werner (2008) and chapters 1 and 2 of Gollier (2004). For an online reference, Levin (2006) proposes an excellent review of the main concepts. Available online at http://web.stanford.edu/~jdlevin/Econ%20202/Uncertainty.pdf. (Accessed on 11/22/2016).

[3] **vNM utility functions can also account for an infinite number of states of the world. In such case, one would have: $U(\tilde{c}) = \int \mu(c) f(c) dc$, where f is the pdf of $\tilde{c}$.**

[4] **Note that $\bar{c}$ can be thought as a degenerated random variable, taking the value $\bar{c}$ with probability 1.**

The parametric form in Eq. 3 implies that the level of risk aversion is independent of the decision maker's financial capacities (hence the name *Constant Absolute* Risk Aversion, CARA). This particular utility function is therefore coherent with the expected behaviour of most public utility services (municipal authorities will not, for instance, gradually adopt a risk-seeking behaviour regarding the protection of citizens if the city's financial well-being improves). **See Appendix B for additional details, proofs, and references for those claims.**

The economic model developed above is applied to the particular context of frequent flooding on the Montmorency watershed. This context is described in greater details in the next section.

3 Context

3.1 Floods on the Montmorency Watershed

Located in southern Québec, Canada, the Montmorency River watershed covers 1150 km^2, most part of which is densely forested. Approximately 30 000 people reside in ~~teh~~**the** basin, concentrated in its southernmost portion. The northern portion of the watershed lays within the Laurentian Wildlife Reserve, where heavy snowfall precipitation is common. **Figure 2 presents the average monthly values for meteorological variables for this watershed.**

Crystalline rock of the Canadian Shield covers most of the watershed, where the retreat of glaciers left till of an average thickness of 1 m. The southernmost part is covered in sandy sediments from the Champlain Sea. Figure 3 shows the geographical location of the watershed as well as the location of the available meteorological stations and streamflow gauges (see section 3.3).

The Montmorency River experiences quasi-annual ice jams during spring melt, which often enhance the magnitude and frequency of floods within vulnerable inhabited areas. The response time of the watershed is rapid (12 hours). The return period of damaging floods is also short. This makes emergency evacuation and flood damage a common occurrence for riverside residents. Table 1 shows return periods and corresponding streamflow values for the Montmorency River (Leclerc and Secretan, 2012). **The table also provides thresholds for streamflow values used for flood mitigation operations (see section 3.2.2)**. Note that these are given for open-water levels, and take neither ice jams nor the increase in water level due to the presence of ice blocks into account.

The behaviour and consequences of ice jams along the Montmorency River have been the focus of previous studies, such as forecasting river ice breakup (Turcotte and Morse, 2015). Risk analysis and technical solutions (Leclerc et al., 2001) have also been studied, but as of yet not implemented.

The river experienced its worst recorded event in 1964, when a heavy rain system melted a late autumn snow cover, resulting in a 1100 m^3/s flow peak. More recently, an ice cover breakup followed by the formation of an ice jam formation further downstream in January 2008 forced the evacuation of 80 households and damaged four houses. In March 2012, an early spring thaw caused by extreme temperatures ~~caused~~ **induced** a flood resulting in the evacuation of 25 households. ~~More recently,~~ **Then,** in April 2014, an ice jam breakup caused a massive ice-carrying flood wave that, occurring during a typical normal spring freshet, quickly raised waters to a semi-centennial level. In addition, the topography in the area causes certain regions

to become entirely isolated and surrounded by water during flooding. The greatest concern of public authorities occurs when people refuse to evacuate, especially in these flood-prone areas.

3.2 Current forecasting and decision-making process

3.2.1 The hydrological model HYDROTEL

HYDROTEL (Fortin et al., 1995) is a spatially distributed, physics-based model developed and maintained by the Institut National de Recherche Scientifique (INRS). It is used operationally by the DEH, and has been implemented in the Montmorency River watershed since 2008 (Rousseau et al., 2008). The model accepts gridded inputs (precipitation, snow cover, temperature) than can be interpolated using a three station average or Thiessen method. Physical features of the catchment (topography, soil type, hydrographic network) are processed by a companion software called PHYSITEL. It divides the watershed in smaller spatial units called RHHU (Relatively Homogeneous Hydrological Units). Each of the RHHU is then assumed to possess homogeneous physical properties. HYDROTEL then performs the computation of vertical and horizontal water flows.

HYDROTEL offers a range of sub-routines for hydrological processes (interpolation of precipitation, evapotranspiration, snow accumulation and melt, etc.). The user chooses the most appropriate sub-routines depending on the available data. For this study, interpolation of observed precipitation was performed using Thiessen's polygons. No radiation data were available, so evapotranspiration was estimated from an empirical temperature-base method (Fortin, 2000; Bisson and Roberge, 1983) and snowmelt was modelled by a mixed degree-day/energy budget approach. The vertical water budget was performed by the sub-routine BV3C **(in French *Bilan Vertical en 3 Couches*)** that divides the soil into three layers of different composition and depths. Both overland and channel routing was performed using the kinematic wave approach (Lighthill and Whitham, 1955). With this setup, which replicates the model setup used operationally by the DEH, HYDROTEL has 27 parameters, but only 10 were calibrated (default values were used for the other parameters). The calibration already performed by the DEH was kept intact. **This calibration was performed using the Shuffle Complex Evolution algorithm of the University of Arizona (SCE-UA, Duan et al., 1994). The objective function to maximize was the Nash-Sutcliffe Efficiency criterion.** In forecasting mode, HYDROTEL is driven by meteorological forecasts, either deterministic or ensemble-based.~~At each time step of a simulation, HYDROTEL updates and save state variables into four separated files: one for surface runoff, one for snow, one for soil moisture and one for streamflow. For instance, the snow file comprises values for five spatially distributed snow-related state variables: snow water equivalent (SWE), snow cover depth, calorific deficit, liquid water content and albedo.~~

In the actual operational setting, data assimilation is performed manually and indirectly: the forecaster modifies precipitation and/or temperature observed during the previous days until the model's simulation is in agreement with the observed streamflow for the actual day. When the model is run with the modified meteorological inputs, state variables are re-computed and should translate into an improvement of the model's description of the hydrological state of the watershed. The choice of applying modifications to temperature or to precipitation depends mostly on the period of the year and associated dominant hydrological that is processed. Thus, during spring freshet, air temperature is the main forcing that acts on the snow melt rate. Solar radiation is not among HYDROTEL's inputs but is rather estimated empirically, in part through air temperature. Therefore, during

this period of year (early March to late May), perturbations are applied on temperature forcing. During the summer and early fall periods, precipitation forcing is the dominant factor for controlling runoff, soil moisture and eventually streamflow. Perturbations are applied primarily on precipitation from approximately June to November.

3.2.2 Flood alerts

The Direction de l'Expertise Hydrique (DEH) is an administrative unit of the Government of Québec created in 2001 with the mandate to manage the water regime of Québec's rivers and provide streamflow forecasts to municipalities. Since 2008, operational five day, three hour time step streamflow forecasts are distributed to municipal water managers. Those forecasts are always obtained using the semi-distributed physics-based hydrologic model HYDROTEL (Fortin et al., 1995). Although HYDROTEL is a deterministic model, the operational forecasts now largely distributed by the DEH are not purely deterministic but are rather accompanied by a 50% confidence interval. This confidence interval is computed from a statistical model derived from the analysis of past deterministic streamflow forecasts errors for 10 watersheds across the province of Québec. A more detailed description of this statistical method is available in Huard (2013).

After receiving a forecast exceeding a pre-determined flood threshold, municipalities can choose to engage in emergency procedures. In the case of the Montmorency watershed, current measures are mostly reactive (road closure, controlled evacuation of citizens, providing emergency shelters and food) rather than preventive (artificial levees, culverts, etc., Leclerc et al., 2001).

Flood thresholds have been adapted from an hydrodynamic study (Leclerc and Secretan, 2012). Threshold numbers have been conservatively rounded down to compensate for the worsening effect of ice in the channel. Table 1 ~~shows~~ **includes** operational threshold levels for the most vulnerable residential area.

3.3 A concurrent flood forecasting framework based on meteorological ensemble forecasts

3.3.1 Meteorological ensemble forecasts

The alternative forecasting framework proposed in this study involves meteorological ensemble forecasts passed on to HYDROTEL. **Precipitation and temperature ensemble forecasts from the Meteorological Service of Canada (MSC) covering the 2011—2014 period are used. For practical reasons, those forecasts were obtained from the ~~Thorpex~~ THORPEX[5] Interactive Great Grand Ensemble (TIGGE) database managed by the European Center for Medium Range Weather Forecasts (ECMWF).** ~~Precipitation and temperature ensemble forecasts from the Meteorological Service of Canada (MSC) were retrieved, covering the 2011—2014 period~~. The forecasting horizon is five days, with a six hour time step. The MSC meteorological ensemble forecasts comprise 20 members. The initial spatial grid of 0.6° was downscaled to a 0.1° grid through simple bi-linear interpolation during data retrieval.

Observations for precipitation and temperature are measured at five ground stations distributed around the watershed (see Figure 3). Hourly measured data was accumulated and averaged over a three hour time step. Snow survey data interpolated on

[5] **THe Observing system Research and Predictability EXperiment. It is a program led by the World Meteorological Organization.**

a 0.1° grid are also available. They were provided for this study by the DEH. The streamflow gauging station at the river outlet provides measurements at a 15 minute interval, corrected for backwater due to ice cover and then averaged over three hour time steps.

3.3.2 Data assimilation and state variable uncertainty

Appropriate data assimilation is crucial for short-term flood forecasting as it allows the model to begin the forecasting period having the best possible estimate for initial conditions. **In a study involving 20 catchments in Quebec,** Thiboult et al. (2016) showed that the uncertainty for initial conditions dominates the other sources of uncertainty for short-term (1-day to 3-day ahead) streamflow forecasts.

In this study, manual data assimilation was performed according to the guidelines by Mamono (2010) and agree with the procedure followed by operational forecasters at the DEH. This assimilation process relies on the assumptions that: (1) model errors are entirely compensated by the model calibration process, (2) streamflow measurements are error-free, and (3) the only remaining source of error affecting state variables is attributable to meteorological inputs (Mamono, 2010). Additive coefficients ~~between -10 and 10 °C~~ were applied to temperature inputs while multiplicative coefficients ~~between 0.1 and 10~~ were applied to precipitation inputs in order to improve the agreement between simulated and observed streamflow series. **Those perturbations were respectively bounded at [-10,10] and [0.1, 10]. Although those minimal and maximal perturbation values are very large, they truly correspond to the rules applied by the DEH operationally. Of course, the goal is to limit perturbations as much as possible. In this study, the multiplicative coefficient applied to precipitation varied between 0.5 and 2.5. Most additive coefficients for temperature varied between -3 and +2.5, with occasional larger coefficients (up to -7 and +7, on three occasions).** Those perturbations of meteorological inputs were applied uniformly onto the basin for fixed periods of time.

The manual data assimilation described above only improves on the "best guess" of the state variables for each time step. To go one step further, additional perturbations were applied around this best guess estimate in order to account for the uncertainty of initial conditions. To do so, a rudimentary Ensemble Kalman Filter (EnKF, Evensen, 2003) was implemented. From the starting point—constituted by manually assimilated precipitation, temperature and streamflow simulation series—random noise is further applied to precipitation and temperature inputs. Additive perturbations are drawn randomly from $U(-8,8)^\circ$ for temperature. For precipitation, both multiplicative ($U(0.5,1.5)$) and additive ($U(0,0.5)$ mm) perturbations are drawn. The inclusion of additive perturbations for precipitation is due to the fact that strong under-captation is suspected for this catchment. Output uncertainty is modelled by a normal distribution centered on observed streamflow with a standard deviation taken as 10% of the observed streamflow. In this study, data assimilation is a necessity rather than a choice and is not at all the primary objective. For this reason, the limits of the above-mentioned distributions were not optimized as in Thiboult and Anctil (2015). Those limits were fixed according to the guidelines in Mamono (2010) and Abaza et al. (2015) and the experience gained during manual data assimilation. Further refinements of the EnKF model is outside the scope of this study.

The Kalman gain K is computed following Mandel (2006)

$$\boldsymbol{K}_t = \boldsymbol{M}_t \boldsymbol{H}^T (\boldsymbol{H} \boldsymbol{M}_t \boldsymbol{H}_t^T + \boldsymbol{O}_t)^{-1} \tag{4}$$

where $\boldsymbol{M}_t$ is the model error covariance matrix computed according to the perturbations defined above and $\boldsymbol{O}_t$ is the covariance of observation noise also computed according to the perturbations drawn from the normal distribution described above. The matrix $\boldsymbol{H}$ relates the state vectors and observations (so called "observation model"). It can be demonstrated through matrix algebra that Eq. 4 amounts to computing the derivative of the analysis error and setting it equal to zero.

Once the Kalman gain is computed, it is used to weight the credibility of the model error $z_t - \boldsymbol{H}\boldsymbol{X}^-$ relative to the a priori estimation of state variables $\boldsymbol{X}^-$ according to Eq. 5. This leads to the updated model states, $\boldsymbol{X}^+$.

$$\boldsymbol{X}_t^+ = \boldsymbol{X}_t^- + \boldsymbol{K}_t (z_t - \boldsymbol{H}\boldsymbol{X}_t^-) \tag{5}$$

The next section adapts the general framework presented in section 2 to the specifics of the Montmorency watershed.

4 Parametrization of the economic model

The preferences of a decision maker with risk-averse preferences represented by a CARA utility function can be represented as follows:

$$U(s) = \sum_m p_m \frac{-1}{A} \exp\Big\{ -A\Big[-d(Q_m) + b\big(d(Q_m), s, w\big) - s \Big] \Big\} \tag{6}$$

Strictly speaking, each ensemble member Q_m has a probability of occurrence p_m, and corresponds to a given damage $d(Q_m)$. In this study, the damage curve is broken down into 12 categories (i.e. $m = 1, ..., 12$). **This choice of 12 categories it is based on a previous hydraulic study by Leclerc and Secretan (2012) to establish inundation maps. They produced 11 maps, for streamflow varying from 550 to 1050 m^3/s with an increment of 50 m^3/s. This increment of 50 m^3/s, is adopted here, but all thresholds were reduced to be in agreement with streamflow values that induced inundations (see also the operational thresholds mentioned in Table 1). The first category represents all the "no flood" category (i.e. below the lowest threshold).**

Then, Q_m represents the streamflow associated with the m^{th} category and p_m becomes the probability associated to this category, inferred from the number of members that fall within it. Given s, the amount of money spent (w days ahead, see section 4.3 below) on flood emergency measures, the resulting gain (or benefit) in terms of damage reduction is given by $b\big(d(Q_m), s, w\big)$.

While Q_m and p_m are derived directly from the ensemble forecast, d, s and $b\big(d(Q_m), s, w\big)$ must be calibrated from other sources of information related to actual operation and decision history. This can be a challenge, but fortunately in the case of

the Montmorency River, a record of citizen evacuations and corresponding spending for the 2014 flood was available. Although incomplete, this record allows us to guide the estimation of d, s and $b\big(d(Q_m), s, w\big)$.

In this context, the cost of the implementing and operating the forecasting system as such is not considered in s. Of course, when the civil security chooses which forecasting system to put in place, they must consider the cost of implementing this particular system. Nevertheless, once the system is in place, its cost should not affect precautionary spending decisions. This also motivated the choice of CARA utility functions, since they do not depend on "wealth" (which would be affected by the cost of performing the forecast)

4.1 Level of risk aversion A

Risk aversion A is an intrinsic characteristic of each person or organization and could be calculated, given the availability of a sufficiently long record of decisions and associated money spending. However, in the present study, A was left free for the following reasons. First, the available data is not sufficient to credibly calibrate A. Second, as one of the goals of this study is to illustrate how risk aversion influences the value of a forecasting system for a particular problem, it is logical to cover a range of possible As, including the risk-neutral $A = 0$ situation. Therefore, A was made to vary from 0 to 0.01. Although these represent relatively small levels of risk aversion **(Babcock et al. (1993))**, ~~it is shown that they lead to qualitative changes in the decision maker's spending decisions,~~ **preliminary tests have shown that, in the context of this paper, these values were sufficient to illustrate a change in the decision maker's spending decisions** and therefore on the economic value of the concurrent forecasting frameworks. Negative values for A were not considered, as they represent a risk-seeking decision maker, unrealistic in the context of flood mitigation.

4.2 Damages d, spending s and damage reduction b

The material damages to houses and property associated with flood events can be estimated using the flow-damage curve established by Leclerc et al. (2001). **This curve is based on a survey regarding the types of houses in the sector: one or two stories, with or without basement, etc. and their value according to the municipal evaluation. The level of submersion for different streamflow values were obtained through hydraulic simulations. The damage is then deduced from this level of submersion using Gompertz' law (Gompertz, 1825).** The damage expressed in dollars rises exponentially with observed streamflow (m^3/s) and range from \$0 to \$375 000.

In this study, the following parametrization of the benefit function is used:

$$b\big(d(Q_m), s, w\big) = \min\left\{\beta_w \cdot s, \psi \cdot \hat{d}(m)\right\} \tag{7}$$

where $d(m) = \psi\hat{d}(m)$, $\hat{d}(m)$ is the flow-damage curve (Leclerc et al., 2001) for the forecast member m, and β_w and ψ are parameters. This particular parametrization assumes that the benefit of spending is linear, until all damages are avoided. It also implies that it is never optimal to spend more than $\max_m\{\psi \cdot \hat{d}(m)\}$, since additional spending brings no additional benefit, for any possible forecast member.

The parameter β_w has been initially calibrated by assuming $\psi = 1$. By comparing the total amount of money spent in 2014 to alleviate flood damages with the damages (in dollars) predicted by the aforementioned damage curve using the observed streamflow, it was found that the calibrated β_w was less than one. This implies that the civil security service would have spend more than the total amount of possible damage.

This therefore implies the existence of intangible benefits associated with having a flood warning system and spending money to mitigate flood effects. According to Lave and Lave (1991) and Carsell et al. (2004), these intangible benefits include but are not limited to: not putting people's health and security at risk, stress reduction for the population, and building a feeling of trust towards the authorities. In the case of the Montmorency River, there has never been any loss of life. However, as mentioned earlier, it may happen that people refuse to leave their residences and become isolated from communicating roads restricting their access to services and medical care. Unfortunately, it is very difficult and probably rather imprudent to associate a definite cost to these intangible benefits such as "reducing stress". In the absence of a better alternative, in this study a multiplying factor ψ was applied to the damage curve to account for those intangible benefits, as suggested in Van Dantzig and Kriens (1960). The parameter ψ was made to vary between 1.5 to 10 and β_w was computed again for each different value of ψ, as the damage curve is modified. The lower limit of ψ was set so that money spent during the flood of 2014 equals the damage predicted by the damage curve. Therefore, in this framework, the damage curve of Leclerc et al. (2001) (i.e. $\hat{d}(m)$) represents mostly the relationship between streamflow and its impact on the lives and well-being of people.

4.3 Warning time and dynamic decision-making

According to **the** US Army Corps of Engineers (1994), **as well as to**Richardson (2000) and Roulin (2007), the costs of emergency measures or benefits thereof are related to warning time w. In particular, Roulin (2007) assumes that early action can reduce the total cost of emergency measures and maximize damage reduction. Carsell et al. (2004) also provide an evaluation of residential content (furniture, food, electric appliances, etc.) that can be protected with a given warning time.

~~That being said, the decision-making process analyzed in this study is a static one. If the analysis of a dynamic decision process would be more realistic, it would also introduce many more questions regarding how the total spending should be distributed among lead times. For instance, depending on their level of confidence regarding the 5-day forecast, a decision maker could decide to launch an evacuation alert and immediately spend all available funds for emergency measures. As stated in Roulin (2007), intuition lends to thinking that preparing in advance for a flood could lead to reduced overall spending compared with waiting until the last minute. This is also discussed in Morss (2010) in her analysis of three case studies of the interactions between flood forecasts, decisions and outcomes. She provides examples of the importance of early actions:~~

~~*Key property- and life-saving decisions are often thought of as taking specific protective action immediately prior to or during an event. However, sometimes key decisions can be less evident and occur during earlier planning stages. For example, in Grand Forks, once officials had decided to expend most of their time, effort, and resources on planning and building primary dikes, they were not able to plan and build secondary dikes fast enough when the flood grew worse than expected. In the Pescadero case, if officials had not decided to position rescue crews and equipment before the flood began, they would not have been able to reach the area.*~~

However, the accuracy of forecasts is inversely ~~proportional~~ **related** to lead time and the decision maker might want to wait for better information before taking a decision.

Those considerations go far beyond the objective of this study, and the formalization of an explicit dynamic decision process is left for further research. In this study, the dynamic nature of the problem is addressed by assuming that the decision maker uses the following myopic decision procedure:

1. At the beginning of each day, the decision maker receives a 5-day forecast.
2. Iteratively, and starting with the *earliest* (5-day) forecast, the decision maker chooses their preferred level of spending. This level of spending is chosen as to maximize Eq. 6.
3. The decision maker is constrained (by external factors such as the availability of materials or labour force) to spend at most a fraction δ of their preferred level of spending s (see below).

The benefits of a spending are assumed to take effect on the day the spending decision is made, up until the forecast date. For example, if a decision maker spends \$1000 on a given Monday, anticipating a flood the following Thursday (i.e. a 4-day forecast), then any damage occurring prior to Thursday is also reduced (by $\beta_w \times$\$1000).

The parameter β_w is divided between lead times according to $[2, 1.75, 1.5, 1.25, 1]\beta_{2014}$, where β_{2014} is calibrated on the spending decisions of 2014 and represents the baseline ratio of gain per dollar invested. The above multiplication therefore assumes that early actions lead to higher gains per dollar spent. This is very similar to the methodology presented in section 4.3 of Roulin (2007), except that only one repartition of β_w is tested here compared to two in Roulin (2007).

If the decision maker is to take successive actions at different lead times according to forecasted streamflow, then the total amount of available money can be spread across lead times. The decision maker can, for instance, spend all the available money two days prior to the event. Or, they can spend half two days prior and the remaining half the day before the flood (1-day). To account for this, five different "spending vectors" were created (Table 2). The values in those spending vectors represent the maximal fraction δ of the preferred level of spending s that can be spent at each lead time. The first three spending vectors represent situations for which there is no limit on the spending than can be made the day before, with spending vector number 3 representing the extreme case where the decision maker *must* wait until the 1-day forecast before spending any money. On the contrary, spending vectors number 4 and 5 represent a fictitious situation in which the decision maker can spend any amount of money at the 5-day horizon, and no spending is allowed the day before (1-day).

It is important to note that due to the myopic decision-making procedure, the decision maker does not take into account the fact that money spreads across lead times when making a decision. This effect alone underestimates the value of early spending. However, the decision maker also does not consider the reduction in uncertainty gained by waiting (which overestimates the value of early spending). In this study, those two effects are assumed to balance each other.

To summarize, the simulation procedure is as follows:

1. **Fix A and ψ**
2. **Given the spending decision of 2014, infer the value of β_{2014} (given the decision model).**

3. **Given A, ψ, β_{2014} and the other model parameters, apply the decision-making procedure described in section 4.3 for each forecast.**

4. **Compute the measures of performance assessment (see section 5).**

5 Performance assessment

5.1 Forecast quality

The three forecasting systems described in sections 3.2 and 3.3 are compared to each other by assessing their respective abilities to forecast observed streamflow values for the 1- to 5-day projections. ~~Firstly, a visual inspection of the forecasted hydrographs is undertaken~~. **This p**erformance assessment **also** involves the well-known Continuous Ranked Probability Score (CRPS, Matheson and Winkler, 1976) and a reliability diagram (Stanski et al., 1989). ~~A visual inspection of corresponding hydrographs is then undertaken.~~

5.2 Evaluating the benefits of forecasts

As described in the introduction, the usefulness of an early flood warning system is in helping the decision maker choose the best spending level s, prior to the event. The value of such system is therefore closely related to the decision maker's ability to affect the outcome through their spending decisions. The benefits of forecasts are therefore evaluated with an explicit concern for the decision maker's preferences.

In order to develop an indicator of the economic benefits of a forecast, it is important to distinguish between the decision maker's *ex-ante* utility (before the uncertainty is resolved, as in Eq. 6) and their *ex-post* utility (the realized level of utility, after the uncertainty is resolved). This is important as spending decisions are based on the *ex-ante* utility, whereas the value of the forecasts are based on the (expected) *ex-post* utility, conditional to spending decisions. Given the spending decision s and the realized state m, the *ex-post* utility of the decision maker is given by:

$$U_m(s_f) = \frac{-1}{A} \exp\Big\{ -A\Big[-d(Q_m) + b\big(d(Q_m, s_f, w) - s_f\Big]\Big\} \tag{8}$$

where s_f is the total amount of money spent, from a decision based on forecasts (f). The value of this *ex-post* utility is dependent, of course, of the realized streamflow values. In order to obtain a sensible evaluation of the decision maker's utility, one must therefore consider the average *ex-post* utility:

$$\mathbb{E}_m U_m(s)$$

where the expectation $\mathbb{E}_m$ is taken with respect to the historical streamflow values. Note that **, strictly speaking,** the history under consideration ~~must~~ **should** be long enough to be representative of the true distribution of streamflow. **On the one hand, it is expected that a longer record will provide a better empirical estimate of the true streamflow distribution. On the other hand, there can also be various sources of non-stationarity affecting the observed streamflow values over time (e.g.**

changing the measurement apparatus, climate change, land-use change, etc). Hence, even with a very long historical record, the true distribution of streamflow cannot be known with certainty. (Note that this also affects measures of quality, such as the CRPS.)

The average *ex-post* utility can be computed for any of the three forecasting systems described in sections 3.2.2 and 3.3 but also for two special cases: perfect forecasts and no forecasts. On one hand, if forecasts were perfect, there would be no missed events and the decision maker would spend only the exact amount of money necessary to obtain the maximum possible protection, as early as time allowed. On the other hand, if no forecasts were available, there would be no decisions to be made and no money to be spent on flood mitigation and protection measures. Therefore, the maximum amount of damage would occur for each flood event.

It is important to note that utility is an ordinal quantity that only represents the preference of a person faced with a decision-making problem, given some information from uncertain forecasts. That is, the utility levels can be compared, but the actual value of the decision maker's utility has no interpretation. Consequently, the utility values computed for the three forecasting systems can be scaled relative to the utility of a perfect forecasting system. This simplifies the interpretation, without imposing any additional restriction.

The hit rate and the overspending index, two standard measures of the economic performance are also presented.

The hit rate, given by Eq. 9, is the ratio of avoided damages when decision-making is based on the forecasting system being evaluated to the damages that would be avoided if the forecasts were perfect (always equal to the observations).

$$Hit\ Rate = \frac{\mathbb{E}_m b\big(d(Q_m), s_f, w\big)}{\mathbb{E}_m b\big(d(Q_m), s_p, w\big)} \tag{9}$$

where s_p is the amount of money that would have been spent if perfect forecasts would have been available. s_f is the total amount of money spent when decisions are based on forecasts, as in Eq. 8. s_p matches exactly the damages corresponding to the observed streamflow, for all time steps.

Overspending is defined as in Eq. 10. It allows for measuring how much the forecasting system being evaluated overspends (in percentage) compared to perfect forecasts. One should aim for the overspending value to be as low as possible.

$$Overspending = \frac{\mathbb{E}_m s_f - \mathbb{E}_m s_p}{\mathbb{E}_m s_p} \tag{10}$$

Results are presented in the next section. ~~A brief description of the simulation procedure for the computation of the economic value of forecasts can be found in Appendix C.~~

6 Results

6.1 Assessment hydrological forecasts relative to observations

Figure 4 displays hydrographs for a two-week period during the spring of 2014. Panels (a), (c) and (e) correspond to 1-day forecasts while panels (b), (d) and (f) correspond to 5-day forecasts. In all cases the time step is three hours. Forecasts along the upper row (a and b) are dressed deterministic forecasts. Forecasts along the middle row are based on meteorological ensemble forecasts without EnKF while forecasts on the bottom row are also based on meteorological forecasts but account for state variables uncertainty through EnKF. This figure shows that for 1-day forecasts, ~~those~~ **forecasts** based on meteorological ensembles ~~and dressed deterministic forecasts~~ **generally** have **low** ~~similar~~ spread. This is expected, as only the forcing uncertainty is accounted for and this uncertainty requires more than one day to be propagated through the hydrological model.**In addition, at short lead times the members of meteorological ensemble forecasts are often very similar. However, before each of the two flood peaks, they display more dispersion than dressed forecasts.** The influence of the EnKF can also be seen. The spread of the forecasts with EnKF is greater than the forecasts without EnKF and the density of forecasts members is higher around the observed streamflow. At the 5-day lead time, some members of the forecasts based on meteorological ensembles reach very high streamflow values. This is not the case for the dressed deterministic forecasts that often underestimate streamflow.

Figure 5 presents the mean CRPS of the three concurrent forecasting systems over the 2011—2014 period. The CRPS was computed separately for each lead time in three hour increments and averaged over the entire record of forecasts and corresponding observations. For very short lead times, the dressed deterministic forecasts outperform those based on meteorological ensembles (lower CRPS).**As noted above, for short lead times the members of the meteorological ensemble forecasts are often very similar and the forecasts thus have no dispersion. Dressed forecasts, by definition, necessarily have more spread. Since the forecasting system is not perfect, an ensemble with very low spread is at risk of missing the observation.** However, for lead times longer than 18 hours, forecasts based on meteorological ensembles achieve a better (lower) CRPS than dressed forecasts, despite the jumpy behaviour of the ensemble curves compared to that of the dressed forecasts. Furthermore, the performance gap between meteorological ensemble-based forecasts and dressed forecasts increases with lead time.

The perturbation of state variables after manual data assimilation increases (worsens) the CRPS. This is likely attributable to a loss of resolution. Although sharpness**,** ~~and~~ resolution and reliability are all desirable attributes of a forecasting system, there is most often a trade-off between the ~~two~~ **resolution** and reliability. **Sharpness is akin to "precision" and refers to the quality of a forecasting system which issue forecast members that are all close together. Resolution is is the ability of the forecasting system to distinguish between different situations.** Indeed, Figure 6 highlights that forecasts based on meteorological ensembles having a perturbation of state variables display a better reliability than when state variables remain unperturbed. The difference is most striking for 1-day forecasts. Figure 6 also shows that dressed deterministic forecasts are more reliable than forecasts based on meteorological ensembles for short lead times (e.g. one day, hollow circles), but less so for longer lead times (e.g. 5-day, hollow triangles). As lead time increases, the accuracy of meteorological forecasts decreases.

However, the spread of forecasts based on meteorological ensembles increases considerably with lead times therefore more often including the observed values at the 5-day lead time compared to the 1-day lead time.

6.2 Assessment of hydrological forecasts in terms of economic value

For each of the simulated values of A and ψ, the application of each spending vector (c.f. Table 2) was tested over the study period (2011-2014). This section describes the simulation procedure. ~~Additional details are found in Appendix C.~~

An example of the applied methodology and corresponding results is provided in Figure 7. The upper row shows 5-day forecasts from the three systems, starting on May 17, 2014. The lower row shows how each member of each forecast is classified into 12 severity classes ranging from non-damaging (class 1) to centennial-scale flooding (class 12) defined after the damage curve.

The utility function (eq. 6) is used successively with the five spending vectors presented in Table 2. The probabilities p_m with $m = 1...12$ in Eq. 6 correspond to the relative frequencies of each category after classification of forecast members that allows for computing the utility as a function of the money spent. The utility curve maximum provides the optimal spending associated with each forecast. Figure 8 illustrates an example for $A = 0.01$ and $\psi = 7$.

Figure 9 presents the utility, hit rate and overspending as a function of parameter ψ for the three flood forecasting systems under study for various levels of risk aversion and for spending vector number 1 (see Table 2). Note that $A = 0$ corresponds to the case of a risk-neutral decision maker. Negative risk aversion values representing risk-seeking behaviour, were not used. As mentioned in section 5.2, any affine transformation of the utility function is admissible. In Figure 9, the utility of a perfect forecast was subtracted from the utility of each concurrent forecasting system and from the "no forecast" situation. This allows the y-axis of the utility plots to start at 0 and provide a common reference. This figure shows that a risk-neutral decision maker prefers having information from forecasts based on meteorological ensembles (with or without EnKF) rather than having no forecasts. However, for higher levels of risk aversion ($A = 0.01$, bottom line of Figure 9), the ~~a decision maker should prefer the "no forecast" situation~~ **forecasting system has no usefulness for low levels of ψ.**

Although this seems counter-intuitive, it can easily be explained by looking at the hydrographs (cf. Figure 4). Forecasts based on meteorological ensembles, in particular using EnKF, have a tendency to generate members with very high streamflow levels. As risk aversion increases, the decision maker puts more weight towards those members, as the associated damage is considerable. This causes the decision maker to spend large amounts of money to “insure” against the potential damage.

As such high streamflow levels are historically rare for the Montmorency River, the decision maker would have been better off not to spend any money and suffer damage during the relatively rare and comparatively small flood events. **The “usual” flood events for the Montmorency River are not as dramatic as what is predicted by the most extreme scenarios of the predictive distribution. However, for a risk averse decision maker, large weights are attributed to those extreme scenarios. This encourages the decision maker to spend large amount of money to mitigate events that in fact never materialize.**

Dressed deterministic forecasts decrease weakly with ψ, relative to the ensemble forecasts. Put differently, for large amounts of material damage, the dressed deterministic forecasts have much higher values than the ensemble forecasts. This is due to the

fact that, for all lead times, ensemble forecasts include members having "unrealistic" streamflow values. This over-forecasting is exacerbated for high values of material damage and a high value of risk aversion. As the concavity of μ increases (due to an increase in the level of risk aversion A), "bad shocks" are weighted more heavily by the decision maker, leading to considerable levels of (over-) spending.

The same effect can be seen for alternative choices of spending vectors. Figure 10 shows the same parameters (utility, hit rate and overspending) as a function of ψ, for the same forecasts, but for spending vector number 2. With this spending vector, the decision maker cannot spend any amount of money five days ahead and can then progressively spend a greater percentage of the available money as the lead time decreases. In such a case, the decision maker should prefer to have access to forecasts based on meteorological ensembles (rather than the no forecast situation) if they are slightly risk-averse ($A = 0.001$). This is explained by the fact that the 5-day forecast (which contains extreme forecast members, c.f. Figure 4) is not used by the decision maker, which limits overspending.

Eventually, a more risk-averse decision maker ($A = 0.01$) should prefer the dressed forecasts over any other forecasting system, for ψ values over 6. This is again attributable mostly to some members of the ensemble systems frequently forecasting flood events that don't materialize. This is confirmed by the overspending graphs on the right-hand side of Figure 10. Hence, in Eq. 6, the optimal level of spending s is less for the dressed forecasts than for the other forecasting systems.

When ψ becomes very large (very important material damages) the utility of the "no forecast" framework decreases rapidly, especially for a more risk-averse decision maker. Then, even if the decision maker generally overspends, all forecasts are preferred to the "no forecast" situation since the damage associated with a flood event are considerable. For high values of ψ, the spending decision effectively acts as an (valuable) insurance policy. The hit rate increases (slightly) with the level of risk aversion. This is expected, as a risk-averse decision maker will attribute more importance to large streamflow values in the ensemble forecast.

The third column of Figure 10 shows that a risk-averse decision maker would reduce their overspending by using a forecasting system based on dressed deterministic forecasts rather than on meteorological ensemble forecasts with or without EnKF. Dressed deterministic forecasts exhibit much less dispersion than EnKF forecasts, which also accounts for state variable uncertainty. As it was remarked earlier, a risk-averse decision maker will put more weight on higher streamflow values in the ensemble. If the spread is large, the ensemble necessarily includes larger streamflow values. It is therefore not surprising that overspending is larger for the ensemble forecast with the larger spread, especially for high values of both A and ψ.

The results for the other spending vectors (c.f. Table 2) are qualitatively similar and are therefore not presented. These results are available as supplementary material.

Figure 11 shows ~~boxplots of $Q_{f,max} - Q_{obs}$~~ **bar graphs of the relative frequency of each class of events, from 2 to 12, for the different forecasting systems and for observations (see section 6.2). The first class, which is the "no damages" class for low streamflow values, is not included. From this figure, it can be seen that all three systems forecast floods more frequently than they should (according to the observed frequencies). This over-forecasting also increases with the forecasting horizon. However, the frequencies computed from the dressed deterministic forecasts (panel a) are closer to the observed frequencies in each class. It can also be noted that the difference between forecasts based on meteorological**

ensembles without EnKF (b) and with EnKF (c) lies in the representation of extreme events at the 1-day lead time. There are more such over-forecasted situations at this lead time when the EnKF is used as part of the forecasting system. This is sufficient for the EnKF forecasts to have lower economic value than the forecasts relying only on meteorological ensembles.

~~for dressed forecasts, forecasts based on meteorological ensembles and forecasts based on meteorological ensembles with state variable uncertainty (EnKF). This graph focuses only on the highest streamflow value of each daily forecast, $Q_{f,max}$. This is the worst-case scenario. It is readily apparent that the worst-case scenario of the dressed forecasts is often lower than the corresponding observed value (i.e. $Q_{f,max} < Q_{obs}$). Dressed deterministic can both over- and under-predict streamflow, with comparable frequencies. The two other forecasting systems most often over-predict streamflow. In addition, when $Q_{f,max} > Q_{obs}$ for dressed forecasts, the magnitude of exceedance is, in most cases, far less than for the two other systems. This implies that the worst-case scenario according to dressed forecasts is usually not as bad as when using meteorological ensembles to issue streamflow forecasts. Our risk-averse decision maker is then less inclined to overspend. Figure 11 shows that such extremes can be *even greater* for forecasts based on meteorological ensembles. Of course, the highest values on Figure 11 (b) and (c) are outliers, meaning that they are exceptions to the general rule. However, even a limited number of largely over-forecasted events can trigger useless spending and therefore decrease the economic value of forecasts. In can also be noted that the difference between forecasts based on meteorological ensembles without EnKF (b) and with EnKF (c) lies in the representation of extreme events at the 1-day lead time. There are more such over-forecasted situations at this lead time when the EnKF is used as part of the forecasting system. This is sufficient for the EnKF forecasts to have lower economic value than the forecasts relying only on meteorological ensembles.~~

7 Discussion

Throughout this paper, the impact of risk-aversion on the economic value of forecasts is assessed for a well-trained end-user. We find that risk-averse end-users mainly consider the less favorable scenarios (upper tail of the predictive distribution in the case of flood forecasting). Thus, although the members of the forecasts are truly equiprobable and presented as such to the end-user, they can still be weighted differently in his or her eyes. This is true for any level of risk aversion, but even more so for high levels of risk aversion. For example, Danhelka (2015) mentions:

The Minister simply asked me what the forecast for Prague was. After I have explained all the known information, forecasts and uncertainties, I gave him my best guess of the peak flow. But his response was: "No, no, no, give me the worst-case scenario; don't tell me numbers you cannot guarantee as not being exceeded".

Therefore, any 'outlier' leads to costly actions and the forecasts become of low or null economic value if these outliers are frequent. A consequence of this is that forecasters may be especially careful about the forecasts for high probability of non-exceedance.

The "real" level of risk aversion for the decision maker for flood emergency measures along the Montmorency River remains unknown due to the insufficient record of decisions and associated spending. However, it can be reasonably

assumed that they are highly risk-averse (Claude Pigeon, personal communications). Considering $A = 0.01$ and Figure 10, the dressed deterministic forecasts provide maximal utility. They have a lower hit rate but also a much lower level of overspending compared to the other forecasting systems. This leads to the conclusion that dressed forecasts have the highest economic value for this level of risk aversion.

However, this conclusion relies on the assumption that benefits are linear. As the level of damage (i.e. $d(m)$) increases, so does the spending needed to alleviate this damage. In a situation where human casualties are possible (resulting in extremely high values of ψ), the spending needs not to increase with the value of the alleviated damages $d(m)$. For example, the cost of an evacuation is not linked to the (somewhat subjective) value associated with human casualties. These considerations are left for further research.

Our study also shows that forecast *quality* (as verified using metrics such as the CRPS) is not always a guarantee of forecast *value* in an economic sense. In this study, the streamflow forecasts based on meteorological ensembles have better CRPS than dressed deterministic forecasts, but their value according to the CARA utility function is lower.

In any case, it is capital to recall that the role of the forecaster is to issue the best possible streamflow forecast given their knowledge of the situation and available model and data. It is the end-user's role to decide the course of action. In no way we would advocate for the forecasters to deliberately bias the forecasts for a certain user. Furthermore, in this paper we did not address the issue of potential cognitive biases and training issues for end-users, which is recognized in the literature (e.g Ramos et al., 2013; Demeritt et al., 2010; Doswell, 2004). The training of end-users and continuous interaction with forecasters should be encouraged to favor optimal decision-making.

Lastly, the decision-making process analyzed in this study is a static one. It would be even more realistic to analyse flood mitigation as a dynamic decision process. For instance, depending on their level of confidence regarding the 5-day forecast, a decision maker could decide to launch an evacuation alert and immediately spend all available funds for emergency measures. As stated in Roulin (2007), intuition lends to thinking that preparing in advance for a flood could lead to reduced overall spending compared with waiting until the last minute. This is also discussed in Morss (2010) in her analysis of three case studies of the interactions between flood forecasts, decisions and outcomes. She provides examples of the importance of early actions:

Key property- and life-saving decisions are often thought of as taking specific protective action immediately prior to or during an event. However, sometimes key decisions can be less evident and occur during earlier planning stages. For example, in Grand Forks, once officials had decided to expend most of their time, effort, and resources on planning and building primary dikes, they were not able to plan and build secondary dikes fast enough when the flood grew worse than expected. In the Pescadero case, if officials had not decided to position rescue crews and equipment before the flood began, they would not have been able to reach the area.

However, the implementation dynamic decision model also introduces many more questions regarding how the total spending should be distributed among lead times. It is thus left for further studies.

8 Conclusions

The purpose of this study is to set the basis of an alternative framework to replace the cost-loss ratio in economic assessment of early warning flood forecasting systems. This alternative framework is based on the Constant Absolute Risk Aversion (CARA) utility function which is well-known in economics. To the authors' knowledge, risk aversion is rarely, if ever, accounted for in hydro-economic assessment of flood warning systems. This new framework is used to compare the economic value of three concurrent streamflow ensemble forecasting systems using the flood-prone Montmorency River watershed in Quebec, Canada. This study concentrates on ensemble rather than deterministic forecasts, as the recent literature clearly states that ensemble forecasts are preferable to deterministic ones for numerous reasons (e.g. Krzysztofowicz, 2001; Jaun et al., 2008; Velazquez et al., 2010; He et al., 2013). Furthermore, real-life operations for the case study involved here (flood forecasting for the Montmorency River) do not involve deterministic forecasts. However, there exists many different means of constructing streamflow ensemble forecasts: dressed deterministic forecasts, single hydrological models fed with meteorological ensemble forecasts, multiple hydrological models, with or without data assimilation, etc. Those different forecasting systems can be compared in terms of their correspondence with the observation an**d** in terms of their value for an end-user.

The importance of the level of risk aversion of the decision maker for the determination of the economic value of a streamflow forecasting system is illustrated by our results. A risk-neutral decision maker, as assumed in the cost-loss ratio framework, is rarely, if ever, encountered in real-life decision problems. The value of forecasting systems strongly depends on the decision maker's level of risk aversion and this parameter should be as much as possible targeted to the end-user. The results also show that forecast quality as assessed by the CRPS, or the reliability diagram, do not necessarily translate directly into a greater economic value, especially if the decision maker is not risk-neutral. Frequent over-forecasting strongly affects the economic value of forecasts. Over-forecasting can be corrected by adequate statistical post-processing of the predictive distributions. This was judged outside of the scope of this study but could certainly be explored in further work. Adequate post-processing would likely improve the value of forecasts.

The decision-making framework presented here can be improved in some ways. Further studies could also include a more detailed, dynamic decision-making process, formally accounting for the forecast horizon. Furthermore, the decision maker could lose confidence in a "bad" forecasting system. The results presented in this paper implicitly assumed that the decision maker's trust of the forecast was absolute. Further studies could include an explicit description of the decision maker's learning about the reliability of a forecast.

9 Appendix A: How the cost-loss ratio implies risk-neutrality

Consider the simple case where the decision maker has two possible choices: $s = 0$ (no action) or $s = 1$ (action). The cost of implementing the action is denoted by $c > 0$. If the adverse event occur (e.g. flood), a damage of $d > 0$ is incurred. Let also b be the damage avoided if an action is taken by the decision maker (assuming $c < b \leq d$). Finally, let p be the probability of the adverse event.

Using the economic model presented in section 2, the vNM utility of the decision maker for each of the possible choices is:

$$U(s=0) = p\mu(-d) + (1-p)\mu(0) \tag{11}$$

$$U(s=1) = p\mu(-d+b-c) + (1-p)\mu(-c) \tag{12}$$

Straightforward algebra shows that an action is optimal (i.e. $U(s=1) \geq U(s=0)$) if, and only if,

$$p \geq \frac{\mu(0) - \mu(-c)}{\mu(0) - \mu(-c) + \mu(-d+b-c) - \mu(-d)} \tag{13}$$

If $\mu(\cdot)$ is concave (the decision maker is risk-averse), this is not equal to the cost-loss ratio. However, if the decision maker is risk-neutral, $\mu(\cdot)$ is linear, so for some $a_1 > 0$ and $a_2 \in \mathbb{R}$: $\mu(0) = a_2$, $\mu(-c) = -a_1 c + a_2$, $\mu(-d) = -a_1 d + a_2$ and $\mu(-d+b-c) = a_1(-d+b-c) + a_2$. Therefore, Eq. 13 reduces to:

$$p \geq \frac{c}{b} \tag{14}$$

If $b = d$ (all damages are avoided), this gives the usual cost-loss ratio.

Here, an important comment is in order. One could always define "cost" and "loss" as follows:

$$cost = \mu(0) - \mu(-c) \tag{15}$$

$$loss = \mu(0) - \mu(-c) + \mu(-d+b-c) - \mu(-d) \tag{16}$$

so an action is optimal if and only if:

$$p \geq \frac{cost}{loss} \tag{17}$$

However, this "black-box" analysis side-steps some interesting and important questions regarding the contribution of outcome versus risk preferences to the decision maker's utility. Using the vNM utility allows us to explicitly describe the impact of risk preferences on the value of forecasting systems. Note also that the hydrological literature (e.g. Roulin, 2007; Verkade and Werner, 2011; Muluye, 2011) almost always considers "cost" and "loss" to be defined in monetary units.

To see more clearly the impact of risk-aversion on the optimal decision, suppose that μ is CARA, i.e. $\mu(x) = \frac{-1}{A}\exp\{-Ax\}$, and that $b = d$. Using the formula above and straightforward algebra, we find that an action is optimal if

$$p \geq \frac{\exp\{Ac\} - 1}{\exp\{Ad\} - 1} \equiv t(A) \tag{18}$$

as opposed to $p \geq c/d$ for the cost-loss ratio. One can verify that $t(A)$ is strictly decreasing with $\lim_{A\to 0} t(A) = c/d$. Then, this implies that, as risk aversion increases, the decision maker requires lower confidence level (for the realisation of the adverse event) in order to take an action. The limiting case, when the decision maker is risk neutral, gives the cost-loss ratio.

10 Appendix B: Properties of the CARA utility function

We have: $\mu(x) = \frac{-1}{A}\exp\{-Ax\}$ for some real values for x and $A \neq 0$. One can easily verify that the first derivative **with respect to** x is: $\mu'(x) = \exp\{-Ax\} > 0$, and that the second derivative **with respect to** x is $-A\exp\{-Ax\}$. Therefore, μ is strictly concave if $A > 0$ and strictly convex if $A < 0$. **Figure 1 illustrates a generic example for a CARA utility function.**

The value of A reflects the decision maker's level of risk aversion. Specifically, the *Arrow-Pratt index of absolute risk aversion* is defined as

$$A(\mu) = \frac{-\mu''(\cdot)}{\mu'(\cdot)} \tag{19}$$

for all twice continuously differentiable function $\mu(\cdot)$. If $A(\mu) > A(\tilde{\mu})$, we say that the decision maker whose preferences are represented by μ is more risk-averse than a decision maker whose preferences are represented by $\tilde{\mu}$.

Using the parametric form: $\mu(x) = \frac{-1}{A}\exp\{-Ax\}$, we immediately see that $A(\mu) = A$. Since $A(\mu)$ is independent of x, we say that μ exhibits a constant absolute level of risk aversion.

Note that the CARA utility functions are only defined for $A \neq 0$. However, since an individual is risk-neutral if and only if μ is linear, the utility function of any risk-neutral individual has the form $\mu(x) = a_1 x + a_2$ for $a_1 > 0$ and $a_2 \in \mathbb{R}$. In other words, there is no need to define a specific class of utility for risk-neutral individuals. As such, the CARA utility class needs only to apply to non-risk-neutral individuals.

The interested reader can consult chapter 2 in Gollier (2004), chapter 6 in Mas-Colell et al. (1995) or Levin (2006) for additional details.

Author contributions. Simon Matte performed all the computation and prepared most figures. He also wrote a preliminary version of some portions of the manuscript. Marie-Amélie Boucher initiated the project and coordinated the work. She did most of the literature review, most of the writing and prepared Figures 1, 2 and 10. Vincent Boucher proposed the economic model, prepared the Appendixes and wrote important portions of the manuscript. Thomas-Charles Fortier-Fillion provided the model and hydro-meteorological data. He participated in the interpretation of results all along the project and reviewed the manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work was funded by a NSERC Discovery grant to Marie-Amélie Boucher. Vincent Boucher gratefully acknowledges financial support from the *Fonds de recherche du Québec – Société et culture* and the Social Sciences and Humanities Research Council. The authors wish to acknowledge Quebec's Direction of Hydrological Expertise for providing hydro-meteorological data and the model used in this study. The authors also thank the ECMWF for the development and maintenance of the TIGGE data portal allowing free access to meteorological ensemble forecasts for research purposes. Finally, this work would not have been possible without the much appreciated collaboration of Mr. Claude Pigeon, responsible for public security for the City of Quebec who, among other things, provided the economic database for the flood of 2014.

References

Abaza, M., Anctil, F., Fortin, V., and Turcotte, R.: A comparison of the Canadian global and regional meteorological ensemble prediction systems for short-term hydrological forecasting, Monthly Weather Review, 142, 2561–2562, 2014.

Abaza, M., Anctil, F., Fortin, V., and Turcotte, R.: Exploration of sequential streamflow assimilation in snow dominated watersheds, Advances in Water Resources, 80, 79–89, 2015.

Babcock, B. A., Choi, E. K., and Feinerman, E.: Risk and probability premiums for CARA utility functions, Journal of Agricultural and Resource Economics, pp. 17–24, 1993.

Beven, K.: Facets of uncertainty: epistemic uncertainty, non- stationarity, likelihood, hypothesis testing, and communication, Hydrological Sciences Journal, 61, 1652–1665, 2016.

Bisson, J.-L. and Roberge, F.: Prévisions des apports naturels: Expérience d'Hydro-Québec, Atelier sur la prévision du débit, Toronto, 1983.

Boucher, M.-A., Tremblay, D., Delorme, L., Perreault, L., and Anctil, F.: Hydro-economic assessment of hydrological forecasting systems, Journal of Hydrology, 416—417, 133–144, 2012.

Carpentier, P.-L., Gendreau, M., and Bastien, F.: Long-term management of a hydroelectric multireservoir system under uncertainty using the progressive hedging algorithm, Water Resources Research, 49, 2812–2827, 2013.

Carsell, K., Pingel, N., and Ford, D.: Quantifying the benefit of a flood warning system, Natural Hazard Review, 5, 131–140, 2004.

Cerdá Tena, E. and Quiroga Gómez, S.: Cost-Loss Decision Models with Risk Aversion, 01, Instituto Complutense de Estudios Internacionales, 2008.

Côte, P. and Leconte, R.: Comparison of stochastic optimization algorithms for hydropower reservoir operation with ensemble streamflow prediction, Journal of Water Resources Planning and Management, 142, doi:10.1061/(ASCE)WR.1943-5452.0000575, 2016.

Danhelka, J.: On decisions under uncertainty, http://hepex.irstea.fr/on-decisions-under-uncertainty/, published online: 2015-05-01, 2015.

Demeritt, D., Nobert, S., Cloke, H., and Pappenberger, F.: Challenges in communicating and using ensembles in operational flood forecasting, Meteorological Applications, 17, 209–222, 2010.

Diomede, T., Nerozzi, F., Paccagnella, T., and Todini, E.: The use of meteorological analogues to account for LAM QPF uncertainty, Hydrology and Earth System Science, 12, 141–157, 2008.

Doswell, C.: Weather forecasting by Humans - Heuristics and Decision Making, Weather and Forecasting, 19, 1115–1126, 2004.

Duan, Q., Sorroshian, S., and Gupta, V.: Optimal use of the SCE-UA global optimization method for calibrating watershed models, Journal of Hydrology, 158, 265–284, 1994.

Evensen, G.: The Ensemble Kalman Filter: theoretical formulation and practical implementation, Ocean Dynamics, 53, 343–367, 2003.

Fishburn, P.: Retrospective on the Utility Theory of von Neumann and Morgenstern, Journal of Risk and Uncertainty, 2, 127–158, 1989.

Fortin, J.-P., Moussa, R., Bocquillon, C., and Villeneuve, J.-P.: HYDROTEL, un modèle hydrologique distribué pouvant bénéficier des données fournies par la télédétection et les systèmes d'information géographique, Revue des Sciences de l'Eau / Journal of Water Science, 8(1), 97–124, 1995.

Fortin, V.: Le modèle météo-apport HSAMI : historique, théorie et application, Rapport de recherche (Révision 1.5), Tech. rep., Institut de Recherche d'Hydro-Québec, 2000.

Franz, K. and Ajami, N.: Hydrologic ensemble prediction experiment focuses on reliable forecasts, Eos, 86, 239, 2005.

Gollier, C.: The economics of risk and time, MIT Press, 2004.

Gompertz, B.: On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies, Philosophical Transactions of the Royal Society of London, 115, 513–583, 1825.

Hamill, T. and Whitaker, J.: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application, Monthly Weather Review, 134, 3209–3229, 2006.

He, Y., Wetterhall, F., Cloke, H., Pappenberger, F., Wilson, M., Freer, J., and McGregor, G.: Tracking the uncertainty in flood alerts driven by grand ensemble weather predictions, Meteorological Applications, 16, 91–101, 2013.

Huard, D.: Analyse et intégration d'un degré de confiance aux prévisions de débits en rivière, Tech. rep., David Huard Solution, Quebec, 2013.

Jaun, S., Ahrens, B., Walser, A., Ewen, T., and Schar, C.: A probabilistic view on the August 2005 floods in the upper Rhine catchment, Natural Hazard and Earth System Sciences, 8, 281–291, 2008.

Juston, J., Kauffeldt, A., Montano, B., Seibert, J., Beven, K., and Westerberg, I.: Smiling in the rain: Seven reasons to be positive about uncertainty in hydrological modelling, Hydrological Processes, 27, 1117–1122, 2013.

Katz, R. and Murphy, A.: Economic value of weather and climate forecasts, Cambridge University Press, New York, 1997.

Krzysztofowicz, R.: Expected utility, benefit, and loss criteria for seasonal water supply planning, Water Resources Research, 22, 303–312, 1986.

Krzysztofowicz, R.: The case for probabilistic forecasting in hydrology, Journal of Hydrology, 249, 2–9, 2001.

Lave, T. and Lave, L.: Public perception of the risks of floods: Implications for communication, Risk Analysis, 11, 255–267, 1991.

Leclerc, M. and Secretan, Y.: Reconstruction de la prise d'eau de l'Arrondissement Charlesbourg – Simulation hydrodynamique du secteur Canteloup, des Îlets, Trois-Saults de la rivière Montmorency, Tech. Rep. R1416, INRS-Eau and Laval University, Quebec, 2012.

Leclerc, M., Morse, M., Francoeur, J., Heniche, M., Boudreau, P., and Secretan, Y.: Analyse de risques d'inondations par embâcles de la rivière Montmorency et identification de solutions techniques innovatrices – Rapport de la Phase I – Préfaisabilité, Tech. Rep. R577, INRS-Eau and Laval University, Quebec, 2001.

Levin, J.: Choice under uncertainty, Lecture Notes, http://web.stanford.edu/%7Ejdlevin/Econ%20202/Uncertainty.pdf, 2006.

Lighthill, M. and Whitham, G.: On kinematic waves, I. Flood movement in long rivers, Proceedings of the Royal Society, Series A, 229, 281–316, 1955.

Mamono, A.: Mise à jour des variables d'état du modèle hydrologique HYDROTEL en fonction des débits mesurés, Master's thesis, Université du Québec à Montréal, 2010.

Mandel, J.: Efficient implementation of the Ensemble Kalman Filter, Tech. Rep. R1416, University of Colorado at Denver and Health Sciences Center, Denver, 2006.

Marty, R., Zin, I., Obled, C., Bontron, G., and Djerboua, A.: Toward real-Time daily PQPF by an analog sorting approach: application to flash-flood catchments, Journal of Applied Meteorology and Climatology, 51, 505–520, 2012.

Mas-Colell, A., Whinston, M. D., and Green, Jerry, R.: Microeconomic theory, vol. 1, Oxford University Press New York, 1995.

Matheson, J. E. and Winkler, R. L.: Scoring rules for continuous probability distributions, Management Science, 22, 1087–1096, 1976.

Merz, B., Elmer, F., and Thieken, A.: Significance of "high probability/low damage" versus "low probability/high damage" flood events, Natural Hazards and Earth System Sciences, 9, 1033–1046, 2009.

Morss, R.: Interactions among flood predictions, decisions, and outcomes: Synthesis of three cases, Natural Hazards Review, 11, 83–96, 2010.

Muluye, G.: Implications of medium-range numerical weather model output in hydrologic applications: Assessment of skill and economic value, Journal of Hydrology, 400, 448–464, 2011.

Murphy, A.: The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation, Monthly Weather Review, 105, 803–816, 1977.

Pope, R. and Just, R.: On testing the structure of risk preferences in agricultural supply analysis, Agricultural Journal of Agricultural Economics, 73, 743–748, 1991.

Ramos, M.-H., van Andel, S., and Pappenberger, F.: Do probabilistic forecasts lead to better decisions?, Hydrology and Earth System Sciences, 17, 2219–2232, 2013.

Richardson, D.: Skill and relative economic value of the ECMWF ensemble prediction system, Quarterly Journal of the Royal Meteorological Society, 126, 649–667, 2000.

Rothschild, M. and Stiglitz, J. E.: Increasing risk: I. A definition, Journal of Economic theory, 2, 225–243, 1970.

Roulin, E.: Skill and relative economic value of medium-range hydrological ensemble predictions, Hydrology and Earth System Sciences, 11, 725–737, 2007.

Rousseau, A., Savary, S., and Konan, B.: Implantation du modèle HYDROTEL sur le bassin de la rivière Montmorency afin de simuler les débits observés et de produire des scénarios de crues du printemps pour l'année 2008, Tech. Rep. R921, INRS-Eau, Quebec, 2008.

Schaake, J. C., Hamill, T. M., Buizza, R., and Clark, M.: The hydrological ensemble prediction experiment, Bulletin of the American Meteorological Society, 88, 1541, 2007.

Shorr, B.: The cost/loss utility ratio, Journal of Applied Meteorology, 5, 801–803, 1966.

Stanski, H., Wilson, L., and Burrows, W.: Survey of common verification methods in meteorology, Tech. Rep. World Weather Watch Technical Report No. 8, WMO/TD No.358, David Huard Solution, Geneva, 1989.

Thiboult, A. and Anctil, F.: On the difficulty to optimally implement the Ensemble Kalman filter: An experiment based on many hydrological models and catchments, Journal of Hydrology, 529, 1147–1160, 2015.

Thiboult, A., Anctil, F., and Boucher, M.-A.: Accounting for three sources of uncertainty in ensemble hydrological forecasting, Hydrology and Earth System Science, 20, doi:10.5194/hess-20-1809-2016, 2016.

Turcotte, B. and Morse, B.: River ice breakup forecast and annual risk distribution in a climate change perspective, in: 18th Workshop on the Hydraulics of Ice Covered Rivers, CGU HS Committee on River Ice Processes and the Environment, Quebec, 2015.

US Army Corps of Engineers: Framework for estimating national economic development benefits and other beneficial effects of flood warning and preparedness systems., Tech. Rep. 94-R-3, US Army Corps of Engineers, Alexandria, Virginia, USA, 1994.

Van Dantzig, D. and Kriens, J.: Het economisch beslissingsprobleem inzake de beveiliging van Nederland tegen stormvloeden, pp. 157–170, Maris, A., De Blocq van Kuffeler, V., Harmsen, W., Jansen, P., Nijhoff, G., Thijsse, J., Verloren van Themaat, R., De Vries, J., Van der Wal, L. (Eds.), 1960.

Velazquez, J., Anctil, F., and Perrin, C.: Performance and reliability of multimodel hydrological ensemble simulations based on seventeen lumped models and a thousand catchments, Hydrology and Earth System Sciences, 14, 2303–2317, 2010.

Verkade, J. and Werner, G.: Estimating the benefits of single value and probability forecasting for flood warning, Hydrology and Earth System Sciences, 15, 3751–3765, 2011.

von Neumann, J. and Morgenstern, O.: Theory of games and economic behavior, vol. 60, Princeton University Press Princeton, 1944.

Werner, J.: risk aversion, in: The New Palgrave Dictionary of Economics, edited by Durlauf, S. N. and Blume, L. E., Palgrave Macmillan, Basingstoke, 2008.

Zhu, Y., Toth, Z., Wobus, R., and Mylne, K.: The economic value of ensemble-based weather forecasts, Bulletin of the American Meteorological Society, 83, 73–83, 2002.

Table 1. Streamflow associated with important return periods and flood mitigation thresholds for the Montmorency River watershed.

Return period (years)	Threshold (m^3/s)	Streamflow
	Surveillance: Close surveillance of river behaviour	350
2		439.0
	Pre-Alert: Warning calls to emergency employees	450
	Alert: Mobilization	500
	Flood: Active evacuation	550
5		569.3
10		655.6
25		764.7
50		845.6
100		925.7
1000		1191.2
10 000		1456.0

Table 2. Maximum fraction of total spending s allowed depending of the forecasting horizon. Each spending vector is identified by an identification number (ID) for further reference.

ID	Maximum fraction of spending allowed				
Number	Day 5	Day 4	Day 3	Day 2	Day 1
"No limit for a 1-day forecast"					
1	1	1	1	1	1
2	0	0.25	0.5	0.75	1
3	0	0	0	0	1
"No limit for a 5-day forecast"					
4	1	0.75	0.5	0.25	0
5	1	0	0	0	0

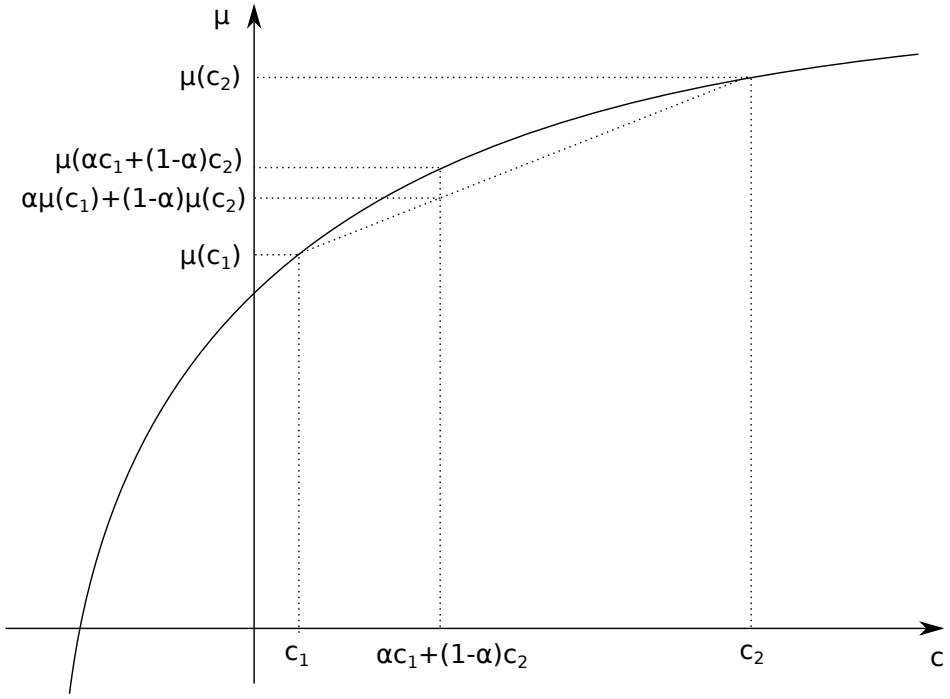

Figure 1. A schematic representation of the CARA utility function for risk-averse individuals. Here, only two states of the world are assumed. The state c_1 is realized with probability α and c_2 is realized with complementary probability. Since μ is concave, we see that the expected utility $U = \alpha\mu(c_1) + (1-\alpha)\mu(c_2)$ is smaller than the utility of the expected value $\mu(\alpha c_1 + (1-\alpha)c_2)$. In other words, the individual would prefer receiving the certain amount $\alpha c_1 + (1-\alpha)c_2$ than receiving a lottery which pays c_1 with probability α and c_2 with probability $1-\alpha$. Equivalently, the individual would be willing to pay *up to* $C = \mu(\alpha c_1 + (1-\alpha)c_2) - [\alpha\mu(c_1) + (1-\alpha)\mu(c_2)] > 0$ to remove the risk associated with this lottery.

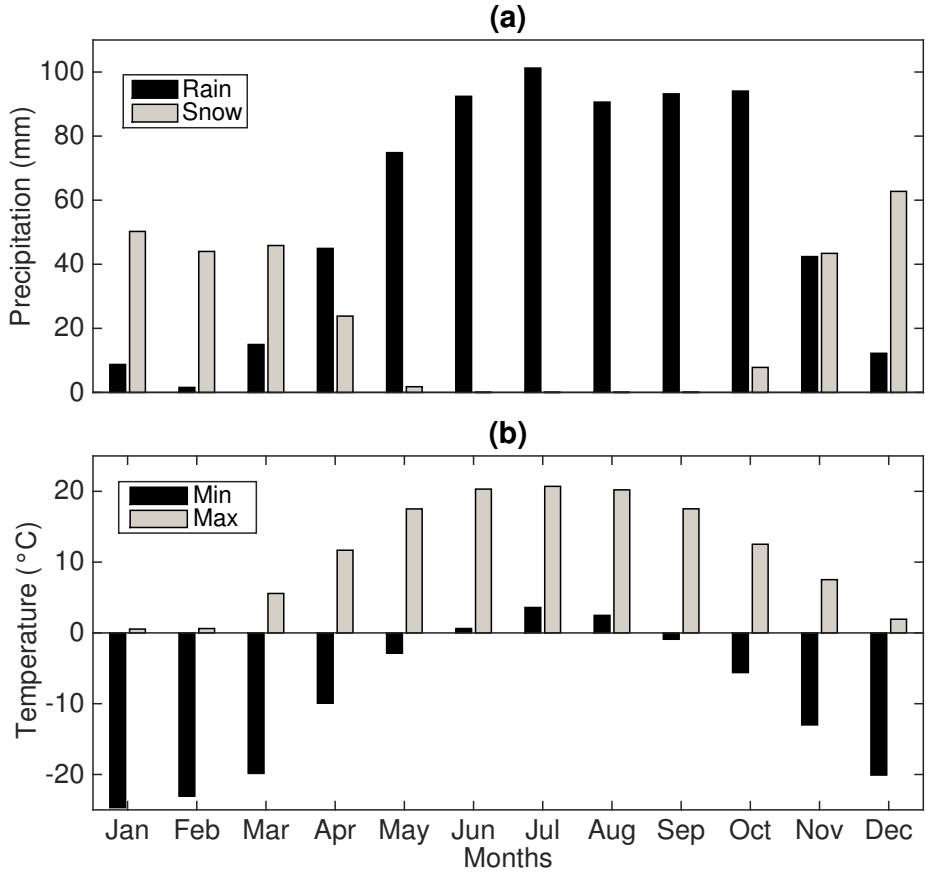

Figure 2. Monthly average values for (a) precipitations and (b) temperature for the Montmorency River watershed

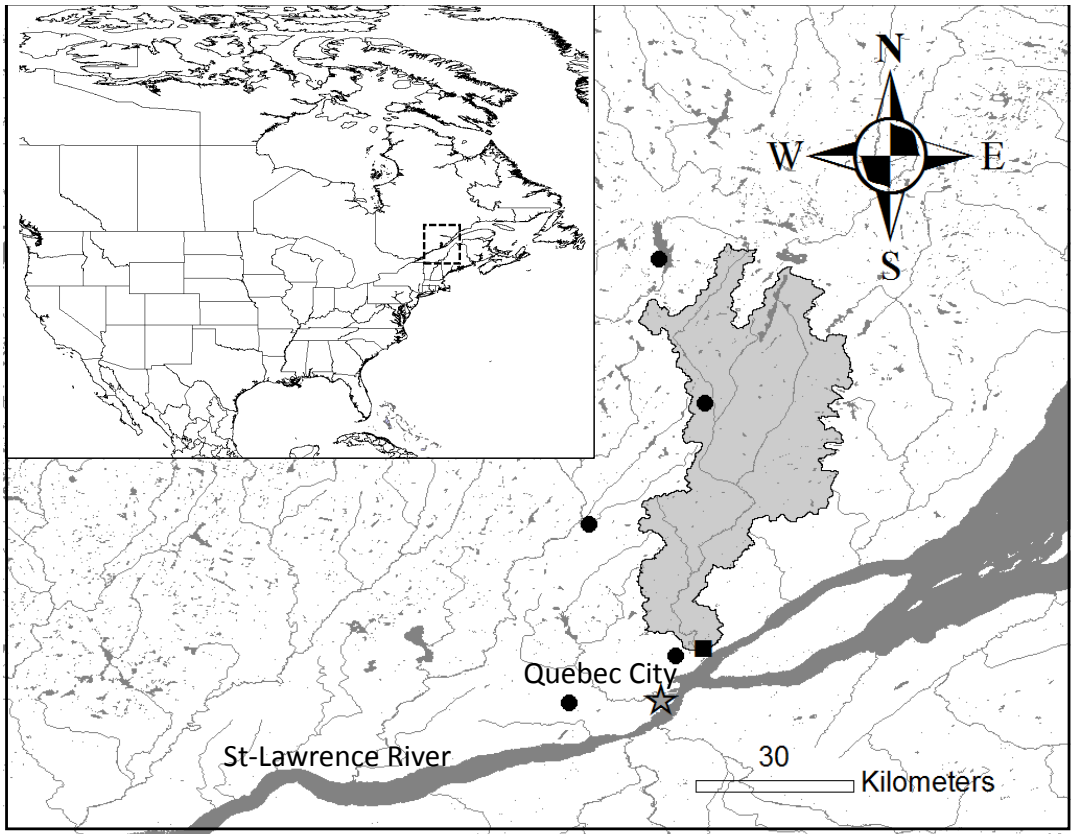

Figure 3. Geographical location of the Montmorency watershed. The black dots represent the available meteorological stations and the black square is the streamflow gauging station.

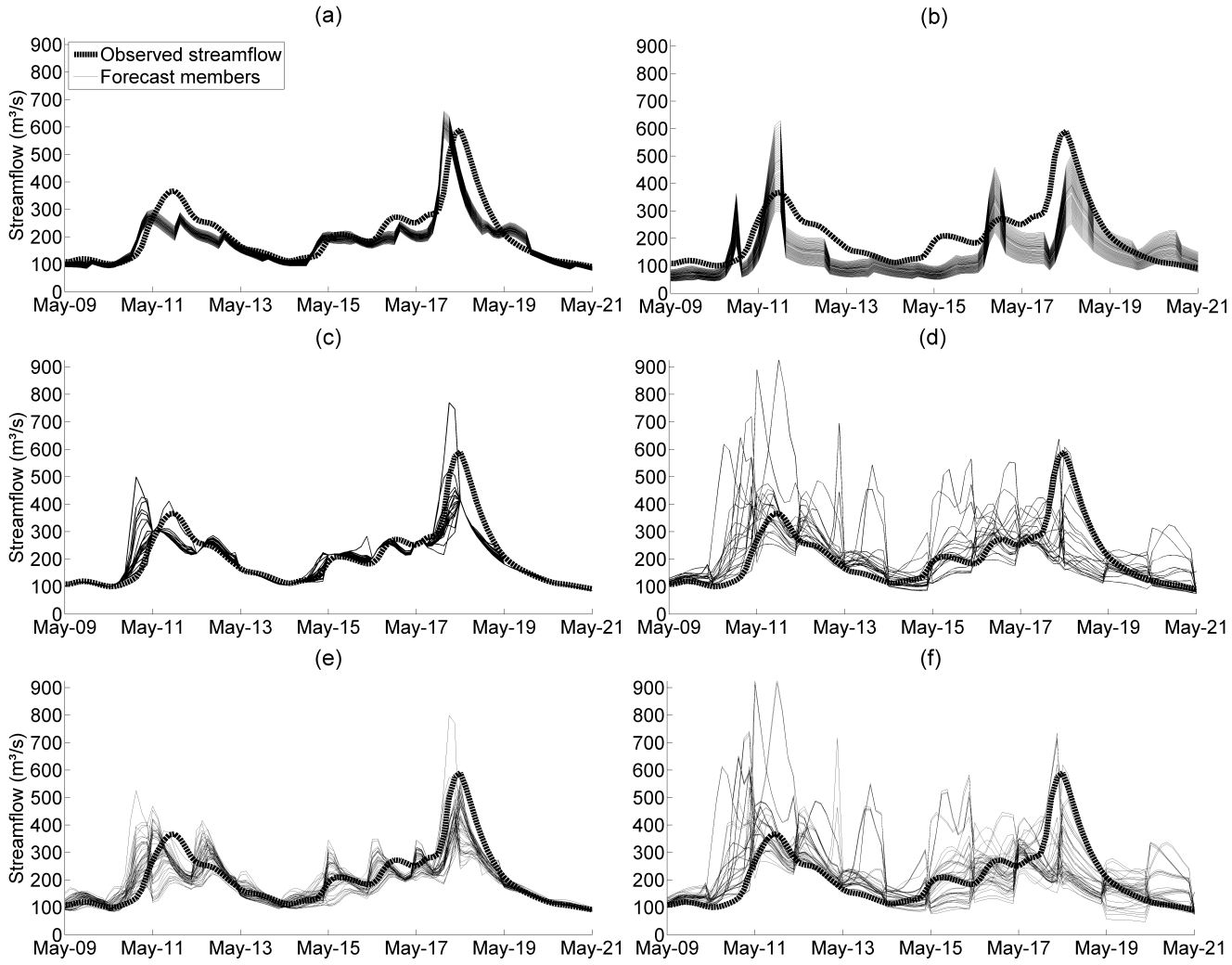

Figure 4. A portion of the 1-day **(left)** and 5-day **(right)** forecasted three hour time step hydrograph in 2014 against the observed streamflow; (a) and (b) are dressed forecasts, (c) and (d) are forecasts based on meteorological ensembles without EnKF and (e) and (f) are forecasts based on meteorological ensembles with state variables uncertainty estimated using the EnKF.

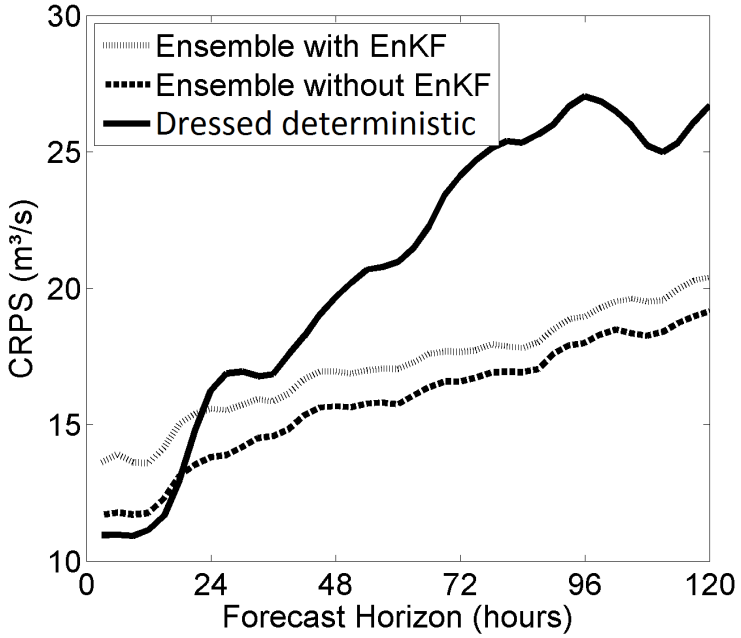

Figure 5. Mean CRPS as a function of lead time for the 2011-2014 period for the forecasts based on meteorological ensembles with (grey line) and without (dashed black line) state variable perturbations and for the dressed forecasts (solid black line).

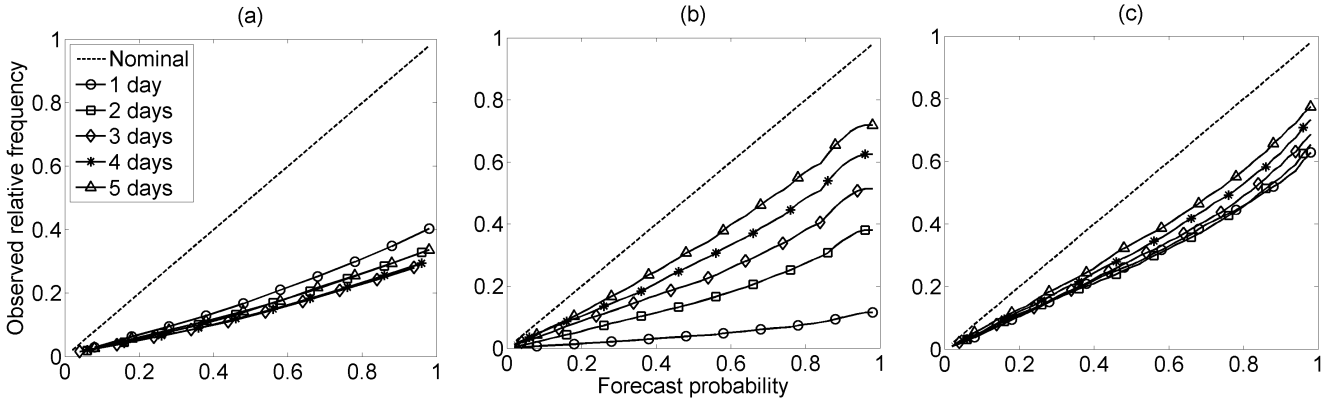

Figure 6. Reliability diagrams as a function of lead time for (a) dressed deterministic forecasts (b) forecasts based on meteorological ensembles and manual data assimilation and (c) forecasts based on meteorological ensembles, manual data assimilation and additional perturbation of state variables.

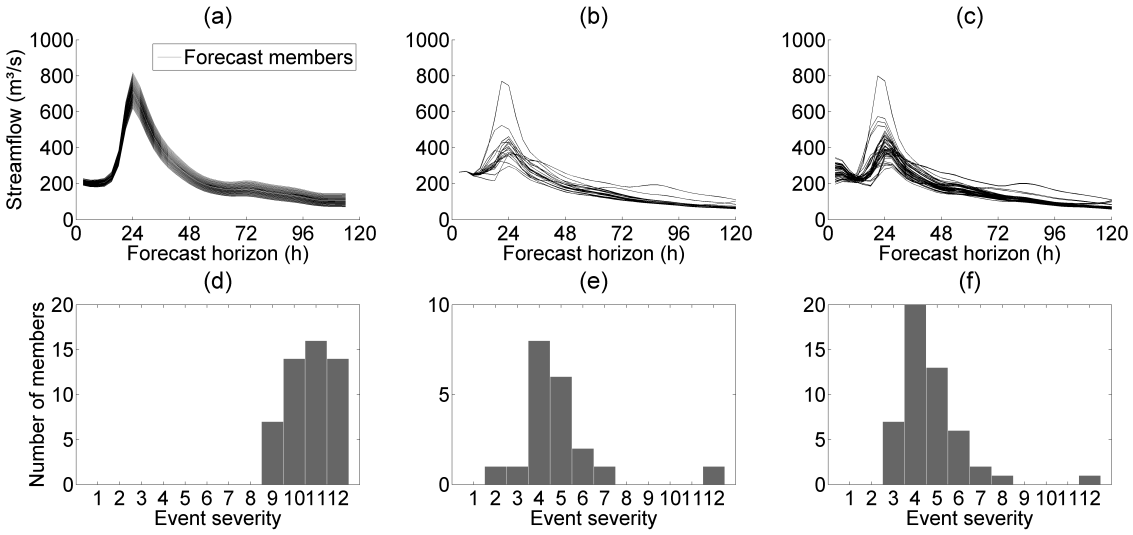

Figure 7. Separation of forecast members into 12 categories according to the magnitude of streamflow. The example is for forecasts emitted on May 17, 2014. (a) and (d) dressed deterministic forecasts, (b) and (e) Meteorological ensemble-based forecasts, (c) and (f) Meteorological ensemble+EnKF forecasts.

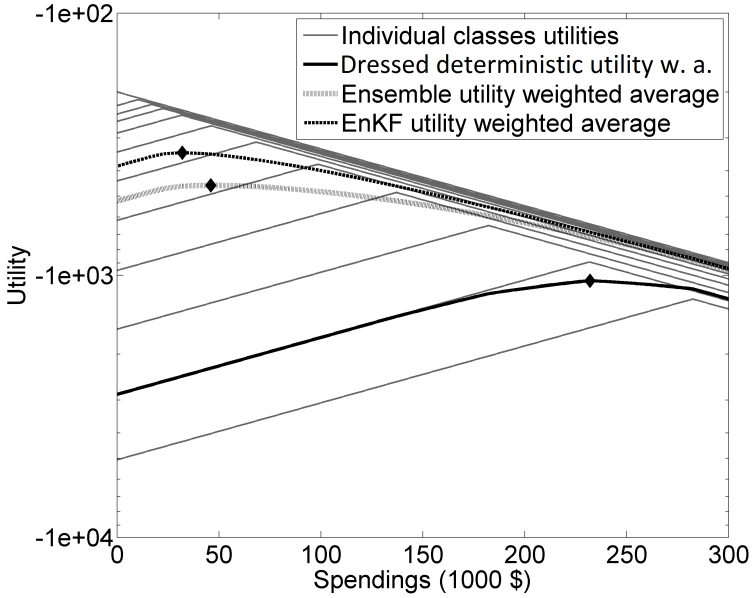

Figure 8. Utility as a function of money spent for forecasts emitted on May 17, 2014 for each of the three forecasting systems. Thin grey curves represent the utility of any decision given the 12 classes of events. Thick curves show the utility of forecasting system. Maxima of each system are indicated by a diamond marker. Calculations are for $A = 0.01$ and $\psi = 7$

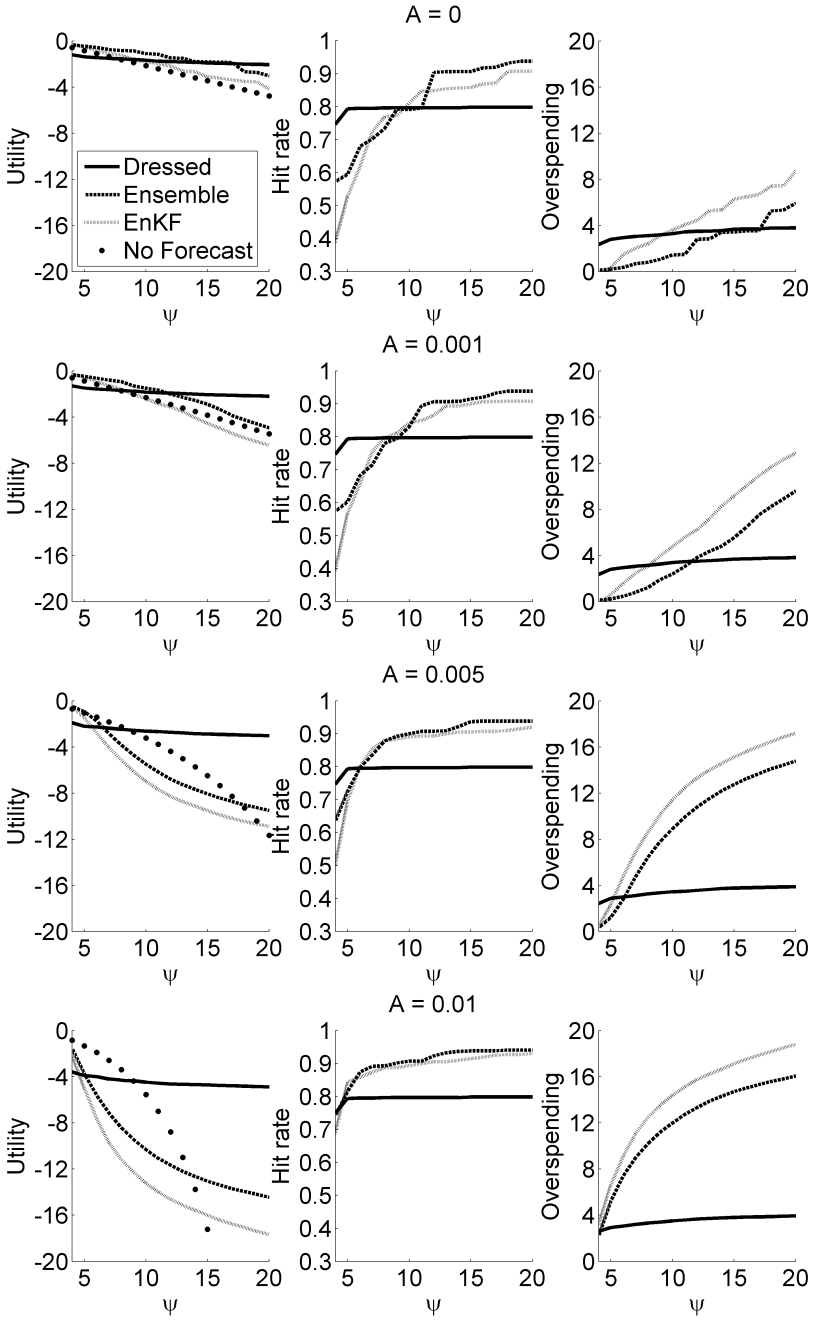

Figure 9. Utility, hit rate and overspending as a function of parameter ψ for the three flood forecasting systems for various levels of risk aversion for the decision maker, when spending is allowed indifferently at any lead time.

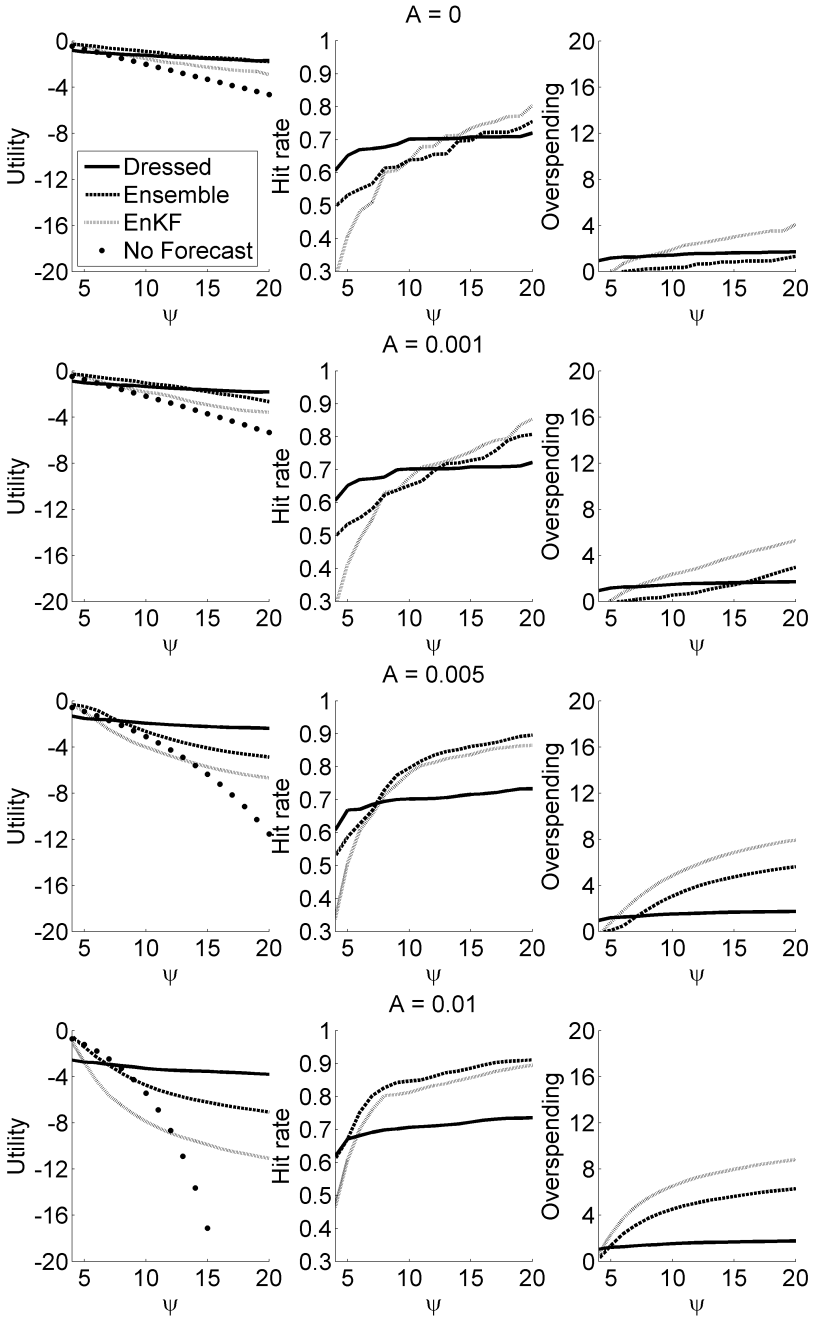

Figure 10. Utility, hit rate and overspending as a function of parameter ψ for the three flood forecasting systems for various levels of risk aversion by the decision maker, when the decision maker is allowed to spend an increasing fraction of the total available money as the lead time shortens.

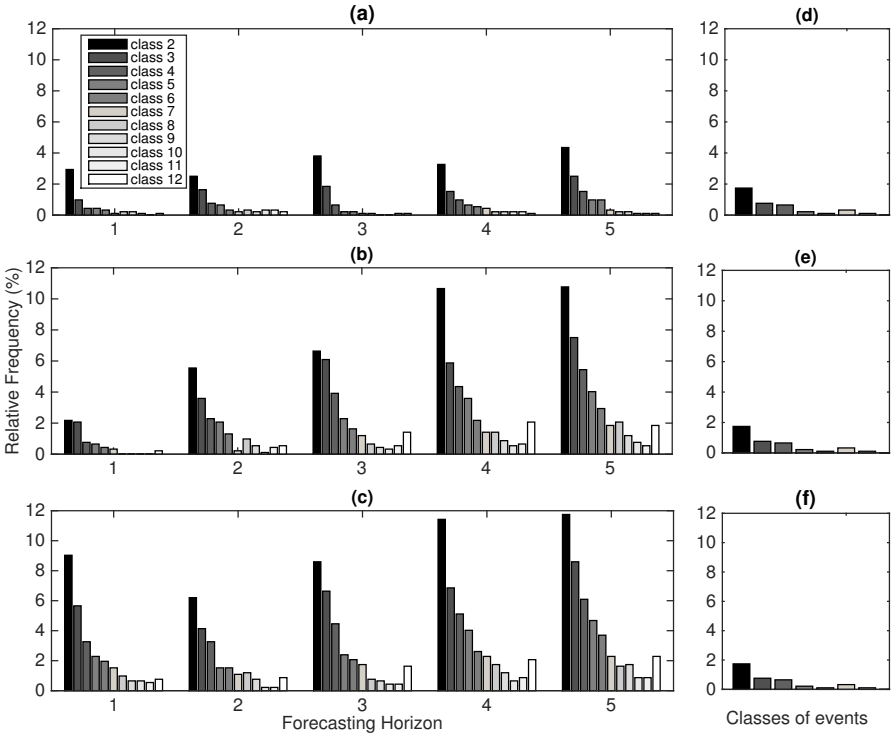

Figure 11. Relative frequencies of forecasts and observations corresponding to the classes of events used in the evaluation of damages, as a function of the forecasting horizon (1 to 5 days). (a) Dressed deterministic forecasts, forecasts based on meteorological ensembles without (b) and with (c) EnKF. Panels (d), (e) and (f) are identical and show the relative frequencies of the observations for the same classes.