Flexible forecast value metric suitable for a wide range of decisions: application using probabilistic subseasonal streamflow forecasts

Richard Laugesen[1,2], Mark Thyer[1], David McInerney[1], Dmitri Kavetski[1]

[1]School of Civil, Environmental and Mining Engineering, University of Adelaide, SA, Australia
[2]Bureau of Meteorology, Canberra, ACT, Australia

Correspondence to: Richard Laugesen (richard.laugesen@bom.gov.au)

Abstract. ~~Forecasts~~Streamflow forecasts have the potential to improve water resource decision-making, but ~~have~~their economic value has not been widely evaluated ~~because~~since current forecast value methods have critical limitations. The ubiquitous measure for forecast value, the Relative Economic Value (REV) metric, is limited to binary decisions, the cost-loss economic model, and risk neutral decision-makers~~.~~ (users). Expected Utility Theory can flexibly model more real-world decisions, but its application in forecasting has been limited and the findings are difficult to compare with those from REV. ~~A~~In this study, a new metric for evaluating forecast value, Relative Utility Value (RUV), is developed using Expected Utility Theory. RUV has the same interpretation as REV, which enables a systematic comparison of results, but RUV is more flexible and ~~able to handle a wider range of~~better represents real-world decisions because ~~all~~more aspects of the decision-context are user-defined. In addition, when specific assumptions are imposed it is shown that REV and RUV are equivalent~~. We demonstrate~~ , hence REV can be considered a special case of the more general RUV. The key differences and similarities between ~~the methods~~REV and RUV are highlighted, with a set of experiments performed to explore the sensitivity of RUV to different decision contexts, such as different decision types (binary, multi-categorical, and continuous-flow decisions), various levels of user risk aversion, and varying the relative expense of mitigation. These experiments use an illustrative case study ~~using~~ of probabilistic subseasonal streamflow forecasts (with lead-times up to 30 days) in a catchment in the southern Murray-Darling Basin of Australia. ~~The ensemble~~The key outcomes of the experiments are (i) choice of decision type has an impact on forecast value – hence it is critically important to match the decision-type with the real-world decision (ii) forecasts ~~were~~ are typically more valuable ~~than a reference climatology~~ for ~~all lead-times (max 30 days), decision types (binary, multi-categorical, and continuous-flow), and levels of risk aversion for most decision-makers. Beyond the second week however, decision-makers who were highly exposed to~~ risk averse users, but the impact varies depending on the decision-context, and (iii) risk aversion impact is mediated by how large the potential damages ~~should use the reference climatology for the~~ are for a given decision. All outcomes were found to critically depend on the relative expense of mitigation (i.e., the cost of action to mitigate damages relative to the magnitude of damages). In particular, for users with relatively high expense of mitigation, using an unrealistic binary decision~~, and forecasts for the~~ to approximate a multi-categorical ~~and~~or continuous-flow decision~~.~~

Style Definition: Appendix: Indent: Left: 0 cm, Hanging: 0.63 cm

Commented [RL1]: 1.5 (and everywhere else user is mentioned)

Commented [RL2]: 2.35 (and other places)

Commented [RL3]: 1.3 (and everywhere else sensitivity analysis is mentioned)

Risk aversion impact was governed by the relationship between the decision thresholds and the damage function, leading to a mixed impact across the different decision types. The generality of RUV makes it applicable to any domain where forecast information is used for making decisions, and gives a misleading measure of forecast value, for forecasts longer than 1 week lead-time. These findings highlight the importance of the flexibility enablesof RUV, which enable evaluation of forecast assessmentvalue to be tailored to specific decisions/users and hence better capture real-world decision-makers. Itmaking. RUV complements forecast verification and enables assessment of forecast systems through the lens of customeruser impact.

Formatted: Font: Not Bold

1 1 Introduction

Formatted: Bullets and Numbering

Effective water resource management is critically important to human welfare, thriving environmental ecosystems, agricultural productivity, power generation, town supply and economic growth (United Nations, 2011; UNESCO, 2012). These decisions depend primarily on the current and anticipated hydrometeorological conditions and are frequently informed by forecasts. In particular, many decisions, such as reservoir operations and early flood warnings, appear to benefit from forecasts at a subseasonal time horizon (2-8 week lead times) because of long river travel times, operational constraints, and logistical overheads (White et al., 2015; Monhart et al., 2019; Schmitt Quedi and Mainardi Fan, 2020; McInerney et al., 2020). These studies use forecast verification techniques and demonstrate that subseasonal streamflow forecasts are becoming more skilful at longer lead times with reliable estimates of uncertainty. However, it is not clear whether forecasts should be used to inform water-sensitive decisions once economic and other factors are considered, the key question is, do the forecasts provide value for the decisions-makers. These factors are typically not considered when evaluating the performance of forecasts using forecast verification. This study aims to address this gap by quantifying the value of subseasonal streamflow forecasts for water-sensitive decisions, such as storage release management and environmental watering.

Forecast verification is the comparison of a set of forecasts spanning a historical period to the observed record using statistical performance metrics. The hydrological forecasting community uses numerous statistical metrics to summarise the performance of ensemble forecasts, including the Continuous Rank Probability Score (CRPS) for accuracy and metrics based on the Probability Integral Transform for statistical reliability for example (Cloke and Pappenberger, 2009; McInerney et al., 2017; Woldemeskel et al., 2018; Bennett et al., 2021). Forecast verification is essential but insufficient for decision-makers to confidently adopt forecasts into their operational and strategic decision-making processes. It does not consider the broader context a decision is made in, the economic trade-offs and different decision types for example. Forecast value measures how much better a decision is when made using one source of forecast information relative to another. It explicitly considers the broader decision context, economics being one of the most tractable aspects to analyse. When using forecast verification metrics as a proxy for forecast value we are implicitly assuming that better verification implies more value. However, additional skill is not necessarily a good predictor of additional benefit to a decision-maker (Murphy, 1993; Roebber and Bosart, 1996; Marzban, 2012).

~~In this paper we consider the value of streamflow forecasts to improve the outcome of binary, multi-categorical, and continuous flow decisions and require a method to quantify this value. However, the most frequently used forecast value method in hydrology and meteorology is Relative Economic Value (REV) which is unable to handle a wide range of decision-types. Substantial research in the field of meteorology has explored the value of temperature, wind and rainfall forecasts for user decisions using REV (e.g. Richardson, 2000; Wilks, 2001; Mylne, 2002; Palmer, 2002; Zhu et al., 2002; Foley and Loveday, 2020; Dorrington et al., 2020). There is an ongoing interest in hydrology to quantify the value of forecasts for decision-making using REV (e.g. Laio and Tamea, 2007; Roulin, 2007; Bergh and Roulin, 2010; Weijs et al., 2010; Verkade and Werner, 2011; Bogner et al., 2012; Abaza et al., 2013; Fundel et al., 2013; Abaza et al., 2014; Thiboult et al., 2017; Verkade et al., 2017; Portele et al., 2021) but no application for subseasonal streamflow forecasts. REV is convenient in its tractability but has strong assumptions about the decision type, economic model, and decision-maker behaviour which neglects important aspects of decision-making and which have implications on the conclusions reached (Tversky and Kahneman, 1992; Katz and Murphy, 1997; Matte et al., 2017).~~

~~REV is only suitable to assess forecast value for a limited set of decisions; binary categorical, cost-loss economic model.~~Effective water resource management is critically important to human welfare, thriving environmental ecosystems, agricultural productivity, power generation, town supply and economic growth (United Nations, 2011; UNESCO, 2012). The management and equitable distribution of water to competing stakeholders is challenging due to long-term decreasing trends in available surface water (Zhang et al., 2016), increasing high intensity storm events (Tabari, 2020), river basins overallocated to irrigated agriculture (Grafton and Wheeler, 2018), and deteriorated river system dependant ecosystems (Cantonati et al., 2020). Environmental decision-making depends largely on the current and anticipated hydrometeorological conditions and is frequently informed by streamflow forecasts. Many decisions, such as reservoir operations and early flood warnings, benefit from forecasts at a subseasonal time horizon (2-8 week lead-times) because of long river travel times, operational constraints, and logistical overheads (White et al., 2015; Monhart et al., 2019). Previous studies have used forecast verification techniques to demonstrate that subseasonal streamflow forecasts are becoming more skilful at longer lead-times with reliable estimates of uncertainty (Schmitt Quedi and Mainardi Fan, 2020; McInerney et al., 2020). However, it is not clear whether forecasts should be used to inform water-sensitive decisions once economic and other factors are considered, thus posing the key question, "do the forecasts provide economic value for decisions-makers?". These factors are typically not considered when evaluating the performance of forecasts, largely due to the limitations of available forecast value methods. This study addresses this gap by developing a new forecast value method that is applicable for a wide range of water-sensitive decisions, such as storage release management and environmental watering.

Forecast verification is the comparison of a set of forecasts spanning a historical period to the observed record using statistical performance metrics. The hydrological forecasting community uses numerous statistical metrics to summarise the performance of ensemble forecasts, including the Continuous Rank Probability Score (CRPS) for accuracy and metrics based on the Probability Integral Transform for statistical reliability (e.g., Cloke and Pappenberger, 2009; McInerney et al., 2017; Woldemeskel et al., 2018; Bennett et al., 2021). Forecast verification is necessary but insufficient for users to confidently adopt

forecasts into their operational and strategic decision-making processes. For example, it does not consider the broader context for which a decision is made, the economic trade-offs and different decision types. Forecast value measures the improvements, in an economic sense, that can be achieved by using one source of forecast information relative to another. It explicitly considers the broader decision context, with economics being one of the most tractable aspects to analyse. When using forecast verification as a proxy for forecast value we are implicitly assuming that better forecast performance (according to our verification metrics) implies more value. However, additional forecast performance is not necessarily a good predictor of additional benefit to a user (Murphy, 1993; Roebber and Bosart, 1996; Marzban, 2012). Exploring the relationship between forecast performance and value over a range of use cases and lead-times is an active area of research, particularly for inflows into hydropower reservoirs (Turner et al., 2017; Anghileri et al., 2019; Peñuela et al., 2020; Cassagnole et al., 2021), and early-warning decision making for extreme events (Bischiniotis et al., 2019; Lopez et al., 2020; Lala et al., 2021).

Commented [RL4]: 2.8

Commented [RL5]: 2.6

Streamflow forecasts can improve the outcomes of a range of decisions, including binary, multi-categorical, and continuous-flow decision types. For example, water level exceeding the height of a levee is a binary decision, and emergency response decisions in relation to a minor, moderate, and major flood classification is a familiar multi-categorical decision. A mitigation decision based on continuous-flow is the limiting case of a very large number of flow classes - for example, adjusting dam releases to match storage inflow during flood operations. While decisions involving more flow classes are an essential feature of many real-world decisions, a binary decision has traditionally been used as the prototypical model of decision making in decision-theoretic literature (Katz and Murphy, 1997). The most frequently used forecast value method in hydrology and meteorology is Relative Economic Value (REV), which is unable to handle a wide range of decision types. Substantial research in the field of meteorology has explored the value of temperature, wind and rainfall forecasts for user decisions using REV (e.g. Richardson, 2000; Wilks, 2001; Mylne, 2002; Palmer, 2002; Zhu et al., 2002; Foley and Loveday, 2020; Dorrington et al., 2020). There is an ongoing interest in hydrology to quantify the value of forecasts for decision-making using REV (e.g., Laio and Tamea, 2007; Roulin, 2007; Abaza et al., 2013; Thiboult et al., 2017; Verkade et al., 2017; Portele et al., 2021), although there have not been applications with subseasonal streamflow forecasts. REV is convenient in its tractability but has strong assumptions about the decision type, economic model, and user behaviour that neglect important aspects of decision-making and have implications on the conclusions reached (Tversky and Kahneman, 1992; Katz and Murphy, 1997; Matte et al., 2017; An-Vo et al., 2019).

Commented [RL6]: 1.26

REV is only suitable to assess forecast value for risk-neutral users making binary decisions using a cost-loss economic model and event frequency ~~as a~~ reference baseline ~~forecast, and risk-neutral decision-makers (Thompson, 1952; Murphy, 1977).~~(Thompson, 1952; Murphy, 1977). This limited setup is an excellent prototypical decision model, which is useful to understand the salient features of forecast value, but may give misleading results when used to model real-world decisions. For example, flood warnings are a practically important multi-categorical decision, typically classified into either minor, moderate, or major flood impact levels, whereas REV only handles binary decisions. Likewise, adjusting the release of water from a storage is best informed by continuous-flow forecasts and may require a more complex economic model than the cost-loss economic model assumed by REV.

REV is also unable to consider the impact of risk-averse ~~decision-makers; a preference for future options~~users. A user is said to be risk averse if they prefer an option with a more ~~certainty~~certain outcome, even ~~though~~if it may on average lead to a less economically beneficial outcome. ~~For example,~~ (Werner, 2008). For example, a water authority deciding to announce a large water allocation event, or an irrigator placing an order ~~for water may prefer forecasts with a more certain outcome (high risk aversion), even if it means missing out on some additional economic benefit. Conversely, a storage operator deciding whether to stop a release because they are concerned about flood damage downstream may,~~ exhibit ~~low aversion to~~ risk ~~and tolerate~~aversion if they prefer a forecast outcome that is almost certain to occur rather than one that is uncertain ~~forecasts of downstream tributary inflows because the economic loss would be significant if a release coincided with a high flow tributary event.~~ but potentially more beneficial.

The field of decision theory explores how agents make decisions with uncertain information and has produced a number of innovations, such as Expected Utility Theory ~~(Neumann, 1944; Mas-Colell, 1995).~~ (Neumann, 1944; Mas-Colell, 1995). Expected Utility Theory is flexible enough to model different decision types, economic models, and risk aversion but there is limited understanding of the relationship and differences between it and REV. It proposes that when faced with a choice a rational person will select the option leading to an outcome that maximises their utility; an ordinal measure based on the ranking of outcomes. Different people may rank outcomes differently because of their specific preferences, such as risk aversion. While Expected Utility Theory is widely used in economics, public policy, and financial management, it has had a very limited application in hydrology and associated fields. ~~Matte et al. (2017) recently used it~~Recently, Matte et al. (2017) applied Expected Utility Theory in a flood damage application to assess the impact of increasing intangible losses and risk aversion on the value of raw probabilistic streamflow forecasts for a single multi-categorical decision type with 12 flow classes. ~~Although the foundation method is general the application~~ This study demonstrated some benefits of forecast value, but was case study specific, limited to a single multi-categorical decision, and used metrics ~~which~~that are somewhat unfamiliar to the verification community. The results were not presented on a traditional Value Diagram and therefore no comparison to REV could be made. ~~The authors~~ We are unaware of any literature ~~which~~that attempts to align REV with forecast value from Expected Utility Theory or present the results on a Value Diagram. There is no method available to the verification community to flexibly evaluate the value of probabilistic forecasts for different decision types, economic models, or ~~decision-makers~~user characteristics ~~(Cloke and Pappenberger, 2009; Soares et al., 2018)~~(Cloke and Pappenberger, 2009; Soares et al., 2018).

~~Probabilistic forecasts of continuous hydrometeorological variables lead to improved forecast verification in many cases and are operationally delivered by all major forecast producers but decision-makers are still learning the most effective way to use them (Duan et al., 2019; Carr et al., 2021). A common approach for decision-makers to use probabilistic forecasts is to first converted them to deterministic forecasts using a fixed critical probability threshold (Fundel et al., 2019; Wu et al., 2020). This approach is known to lead to sub-optimal forecast value in some situations through studies using REV (Richardson, 2000; Wilks, 2001; Zhu et al., 2002; Roulin, 2007). Matte et al. (2017) quantified forecast value with an alternative decision making approach which uses the whole forecast distribution to decide on an ideal action at each forecast update. It is not clear that this alternative approach leads to better decision outcomes and the authors are unaware of any literature comparing them.~~

Commented [RL7]: 1.6

Commented [RL8]: 2.9

Commented [RL9]: 2.10

Probabilistic forecasts of continuous hydrometeorological variables lead to improved forecast performance in many cases and are operationally delivered by all major forecast producers, but users are still learning the most effective way to use them (Duan et al., 2019; Carr et al., 2021). A common approach for decision-making with a probabilistic forecast is to convert it to a deterministic forecast using a fixed critical probability threshold (Fundel et al., 2019; Wu et al., 2020). This approach is known to lead to sub-optimal forecast value in some situations through studies using REV (Richardson, 2000; Wilks, 2001; Zhu et al., 2002; Roulin, 2007). Matte et al. (2017) quantified forecast value with an alternative decision making approach which uses the whole forecast distribution to decide on an ideal action at each forecast update. It is not clear that this alternative approach leads to better decision outcomes and we are unaware of any literature comparing them.

Commented [RL10]: 1.7

Commented [RL11]: 1.8

Commented [RL12]: 2.10

This study aims to:

1. Develop a methodology to systematically compare two forecast value techniques; REV and a method based on Expected Utility Theory.
2. Demonstrate the key differences and similarities between the approaches for different decision types and levels of risk aversion using subseasonal streamflow forecasts in the Murray-Darling Basin.

In Sect. 2, the theoretical background of REV and an Expected Utility Theory approach for forecast value are introduced. Sect. ~~Section~~ 3 proposes a new metric (Relative Utility Value) based on Expected Utility Theory and details its equivalence to REV when a set of assumptions are imposed. Sect. ~~The methodology for a~~4 introduces an illustrative case study using subseasonal forecasts ~~with binary, multi-categorical~~, and ~~continuous flow decisions is introduced in~~ a series of experiments to explore the sensitivity of forecast value to different aspects of decision context. ~~Sect. 4~~. Results of the case study are presented in Sect. 5 and discussed in Sect. 6, including implications for forecast users and producers. Conclusions are drawn in Sect. 7 .

~~2~~ 2 Theoretical background

The background theory introduced here focuses on two methods to quantify the value of forecasts, namely REV and an approach using Expected Utility Theory introduced by Matte et al. ~~(2017)~~(2017).

2.1 Relative economic value

REV is a frequently used and excellent method to quantify the value of forecasts for cost-loss binary decision problems ~~(Richardson, 2000; Wilks, 2001; Zhu et al., 2002).~~(Richardson, 2000; Wilks, 2001; Zhu et al., 2002). Cost-loss is a well-studied economic model where some of the loss due to a future event can be avoided by deciding to pay for an action which will mitigate ~~it~~the loss (Thompson, 1952; Murphy, 1977; Katz and Murphy, 1997). Many real-world decisions, such as insurance, can be simplified and framed in this way as a binary categorical decision. The method assumes that any real-world decision it is applied to can be framed in this way.

2.1.1 REV with deterministic forecasts

Whether a user is expected to benefit in the long run from the use of a forecast system (or an alternative) can be assessed using a ~~4x4~~ 2x2 contingency table. Table 1 includes the hit rate ~~h~~ h, miss rate ~~m~~ m, false alarm rate ~~f~~ f, and correct rejection rate (quiets) ~~q~~ q from a long run historical simulation, along with the net expense from each ~~outcome~~ combination of action and occurrence, where ~~C~~ is the cost of an action to mitigate the loss ~~L~~ L. However, only a portion ~~L_a~~ L_a of the total loss can be avoided with the remainder ~~L_u~~ L_u being unavoidable. A derivation of Eq. (2) is provided in the Supplement.

Table 1: Contingency table for the cost-loss decision problem with expenses from each possible ~~outcome~~combination of action and occurrence. Here ~~C~~ C is the cost of the mitigating action, ~~L_u~~ L_u is the unavoidable portion of loss ~~L~~ L from the event occurring, and ~~L_a~~ L_a is avoidable portion of loss from the action.

	Event occurred	Event did not occur
Action taken	Hit rate ~~(h)~~(h) ~~$C + L_u$~~ $C + L_u$	False alarm rate ~~(f)~~(f) ~~C~~ C
Action not taken	Miss rate ~~(m)~~(m) ~~$L = L_a + L_u$~~ $L = L_a + L_u$	Quiets/correct rejection rate ~~(q)~~(q) ~~0~~ 0

The expected long run expense ~~E~~ E of each ~~outcome~~combination of action and occurrence depends on the rate that ~~outcome~~combination occurred over some historical period, and these rates will be different depending on which forecast

Formatted: Bullets and Numbering
Field Code Changed
Commented [RL13]: 1.9
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Commented [RL14]: 1.10 (and everywhere else outcome is mentioned)
Field Code Changed
Field Code Changed
Commented [RL15]: 1.10
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed

information is used. The REV metric is constructed by comparing the relative difference in the total net expenses for decisions made using forecast, perfect, and climatological baseline information.

~~$V = \frac{E_{climate} - E_{forecast}}{E_{climate} - E_{perfect}}$~~ $\mathrm{REV} = \frac{E_{climate} - E_{forecast}}{E_{climate} - E_{perfect}}$ (1)

~~where each expense term is the summation of the contingency table elements each weighted by the rate of occurrence. Equation (1) is equivalent to the following standard analytical equation for REV (Zhu et al., 2002) when the long run average expenses from Table 1 are considered.~~

~~$V = \frac{\min(\bar{o},\alpha) - (h+f)\alpha - m}{\min(\bar{o},\alpha) - \bar{o}\alpha}$~~ where $-\infty < \mathrm{REV} \leq 1$ and each expense term is the summation of the contingency table elements each weighted by the rate of occurrence. Equation (1) is equivalent to the following standard analytical equation for REV (Zhu et al., 2002) when the long run average expenses from Table 1 are considered.

$$\mathrm{REV} = \frac{\min(\bar{o},\alpha) - (h+f)\alpha - m}{\min(\bar{o},\alpha) - \bar{o}\alpha} \quad (2)$$

~~Where~~ where $\bar{o}$ is the frequency of the binary decision event and the parameter ~~α~~ α is known as the cost-loss ratio.

~~$\alpha = \frac{C}{L_a}$~~ $\alpha = \frac{C}{L_a}$ (3)

~~The derivation of Eq. (2) is available in the Supplement.~~ Equation (2) is typically applied over a range of ~~α~~ α values and this set of REV results is plotted on a Value Diagram. This diagram provides a visualisation of how forecast value varies for ~~decision-makers~~users with different ~~exposure~~levels of costs required to ~~losses~~mitigate a loss, and by extension ~~exposure to~~ mitigation of the underlying damages. ~~We can consider users with a small~~An alternative interpretation of α, which we refer to as *relative-expense of mitigation* (see Table 2), is the relative expense (i.e., cost ~~loss ratio~~) a user experiences to ~~have a smaller~~ take action and mitigate (i.e., avoid) their exposure to damages (i.e., loss). It is a 'relative' expense of mitigation, because the expense magnitude (i.e., cost) is relative to the magnitude of the damages (i.e., loss). This interpretation is used in this study since it is more generalisable across different forecast value methods. Users with smaller cost-loss ratio have a relatively lower expense of mitigation due to their ~~enhanced~~ ability to leverage a ~~small~~lower amount of spending (small cost) to avoid larger future damages (large loss). Conversely, users with a large cost-loss ratio ~~would~~ have ~~a larger exposure to damages.~~relatively high expense of mitigation, as they require a higher amount of spending to avoid future damages. For the same event the ~~level~~relative expense of ~~exposure~~mitigation will vary for different ~~decision-makers~~users and decision types. This *relative-expense of mitigation* interpretation of α should not be confused with the *expected long run expense* E used in the derivation Eq. (2).

Field Code Changed

Commented [RL16]: 1.11

Field Code Changed

Field Code Changed

Commented [RL17]: 1.12

Field Code Changed

Field Code Changed

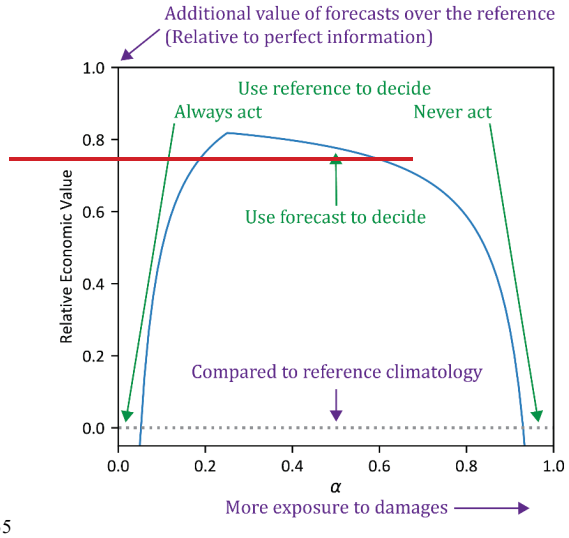
Field Code Changed

Commented [RL18]: 1.10

Field Code Changed

Field Code Changed

Field Code Changed


Additional value of forecasts over the reference
(Relative to perfect information)
Use reference to decide
Always act
Never act
Use forecast to decide
Compared to reference climatology
Relative Economic Value
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
α
More exposure to damages

235

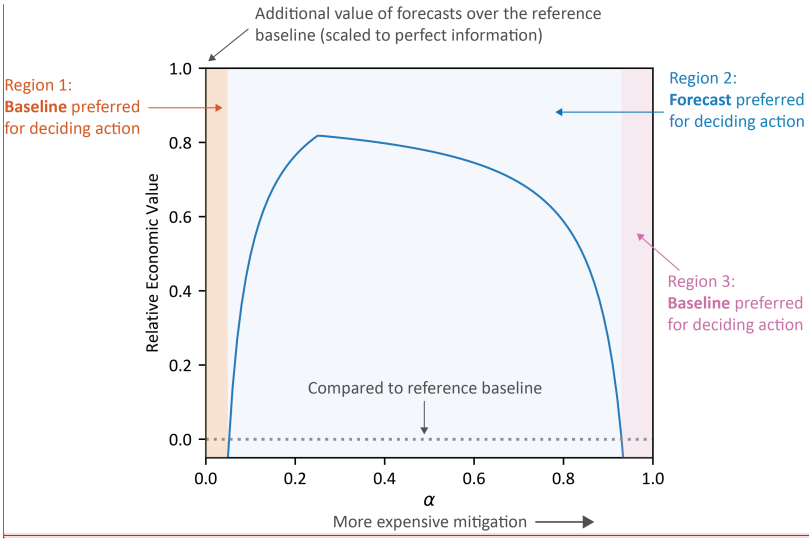

Additional value of forecasts over the reference
baseline (scaled to perfect information)
Region 1:
Baseline preferred
for deciding action
Region 2:
Forecast preferred
for deciding action
Region 3:
Baseline preferred
for deciding action
Compared to reference baseline
Relative Economic Value
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
α
More expensive mitigation

Figure 1: Illustrative value diagram with key features annotated~~. with 3 key regions of α noted. Positive REV for users in region 2 indicates the forecasts should be preferred to the baseline when making the decision under analysis. Negative value for users in region 1 and 3 indicates the reference should be preferred.~~

Field Code Changed

Commented [RL19]: 1.13 and 1.14

Figure 1 presents an illustrative Value Diagram as an aid to describe its interpretation~~. The non-dimensional cost-loss ratio α~~ (Richardson, 2000). The non-dimensional cost-loss ratio α is shown on the x-axis and can be interpreted as a continuum of different decision-makers using the forecasts, with ~~increasing exposure to the damages.~~increasingly more expensive mitigation. A value of ~~$\alpha = 1$~~ $\alpha = 1$ corresponds to maximum ~~exposure:~~relative expense of mitigation; if losses are $100,000 then the amount to spend on a mitigating action is also $100,000. A value of ~~$\alpha = 0.1$~~ $\alpha = 0.1$ indicates that only $10,000 would be needed to mitigate the loss. The y-axis shows forecast value according to REV and has a similar interpretation to any skill-score based metric. A value of ~~REV = 1~~ REV = 1 indicates that decisions made using forecast information successfully mitigated the same level of losses (over the historical period) as decisions made using perfect information (streamflow observations). A value of ~~REV = 0~~ REV = 0 indicates the decisions were only as good as those made using reference ~~forecast information (climatology).~~baseline. A negative value indicates the decisions were worse than the reference. ~~A~~For example, a value of ~~REV = 0.7 , for example,~~ REV = 0.7 at some value of α , indicates that ~~on average the~~ decisions made using forecasts ~~would have successfully mitigated~~led to a 70% ~~more losses over the historical period than~~improvement in net expense relative to decisions made using the reference ~~forecast.~~baseline, a similar interpretation to skill-scores (Wilks, 1995).

Commented [RL20]: 1.15

Formatted: Font: Italic

Field Code Changed

Field Code Changed

Field Code Changed

Formatted: Font: Italic

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

2.1.2 REV with probabilistic forecasts

Constructing a value diagram using Eq. (2) is only possible with ~~categorical~~binary forecasts, so an additional step is required to convert probabilistic forecasts into categorical forecasts to quantify their value.

1. Introduce a critical probability threshold ~~p_t~~ p_τ to convert the probabilistic forecast into a deterministic forecast using the quantile function,
2. Construct a categorical forecast and contingency table from this deterministic forecast and apply Eq. (2) over a range of ~~α~~ α as before,
3. Repeat step 1 and 2 for many probability thresholds over the range ~~$0 \le p_t \le 1$~~ $0 \le p_\tau \le 1$ to form a set of possible REV values for each value of ~~α~~ α ,
4. Take the maximum value from this set for each value of ~~α~~ α to construct a single curve ~~which~~that envelopes the many curves from each value of ~~p_t~~ p_τ ,
5. This envelope is then considered to represent the value of the forecast system.

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Constructing an envelope to represent the forecast value of the system in step 4 can lead to a problematic interpretation. It implicitly assumes that the user will always self-calibrate to select the best critical threshold ~~p_t~~ p_τ for their decision before the event has occurred. This is impractical and the method therefore leads to an over estimation of the expected forecast value. This envelope could ~~be~~ alternatively be interpreted as the maximum attainable forecast value. The impracticality of this method is well understood ~~(Zhu et al., 2002)~~(Zhu et al., 2002) but frequently ignored when applied in practice.

Step 1 of the approach models how ~~decision-makers~~users commonly make decisions using probabilistic forecasts. That is, before the event has occurred (ex ante) a ~~decision-maker~~user will choose a probability threshold that represents the degree of certainty they require to act. If the forecast probability of the event occurring is larger than this threshold then they will act. We refer to this as the *threshold-approach*.

~~Alternatively, one could set the critical probability threshold equal to α which assumes that decision-makers will self-calibrate based on an awareness of their specific exposure to damages (Richardson, 2000). When forecasts are perfectly reliable this approach is equivalent to the maximum forecast value from step 4 (Murphy, 1977). Forecast systems are not perfectly reliable however, even with contemporary post-processing methods (Li et al., 2016b; Woldemeskel et al., 2018; McInerney et al., 2020). The realised value curve will therefore lie below the maximum value curve when applied to real-world forecasts. This alternative approach does not appear to be commonly used by the water resource community to make decisions, or the verification community to assess them.~~

Alternatively, one could set the critical probability threshold equal to α. This approach assumes that the user will self-calibrate based on an awareness of their specific α value (Richardson, 2000). When forecasts are perfectly reliable this approach is equivalent to the maximum forecast value from step 4 (Murphy, 1977). Forecast systems are not perfectly reliable however, even with contemporary post-processing methods (Li et al., 2016b; Woldemeskel et al., 2018; McInerney et al., 2020). The realised value curve will therefore lie below the maximum value curve when applied to real-world forecasts. To the best of our knowledge, studies of real-world decisions using this alternative approach ($p_\tau = \alpha$) have not been reported in the published literature.

Field Code Changed

Field Code Changed

Field Code Changed

Commented [RL21]: 2.11

2.2 Expected Utility Theory approach

Matte et al. ~~(2017)~~(2017) introduced a method to quantify forecast value based on expected utility maximisation with a state dependent utility. The method is flexible enough to model binary, multi-categorical, and continuous-value decisions, along with risk averse ~~decision-makers.~~users. The method assumes that decisions of how much to spend on mitigating damages are based on the forecast probability that the event will occur. We will refer to this approach to decision-making as the *optimisation-approach* to contrast it with the *threshold-approach*.

For a general decision problem with multiple possible future states of world, the following equation specifies the von Neumann-Morgenstern expected utility U for a single timestep t over ~~M~~ M states.

Field Code Changed

Field Code Changed

Field Code Changed

~~$U(\tilde{E}_t) = \sum_{m=1}^{M} p_t^m \mu(E_t^m)$~~ $U_t(E_t) = \sum_{m=1}^{M} p_{t,m} \mu(E_t(m))$ (4)

where ~~p_t^m~~ $p_{t,m}$ is the probability of state ~~m~~ m occurring in timestep ~~t~~ t and ~~E_t^m~~ $E_t(m)$ is the outcome associated with that state. The outcome is typically but not necessarily in monetary units. A utility function ~~$\mu(\cdot)$~~ $\mu(\cdot)$ maps the outcome to a utility. This utility represents an ordinal value that the ~~decision-maker~~user gains from that outcome occurring. The expected utility ~~$U(\tilde{E}_t)$~~ can be considered a probability weighting of the transformed outcomes of all possible states of the world.

Risk aversion is represented by the concavity of ~~$\mu(\cdot)$~~ $\mu(\cdot)$, such that when a ~~decision-maker~~user is risk averse the utility gained from an extra dollar is less than the utility lost when losing a dollar ~~(Mas-Colell, 1995).~~(Mas-Colell, 1995); see Figure 3b for examples of $\mu(\cdot)$ used in our experiments with different levels of risk aversion. Therefore, on average the risk is only worth taking when the probability of gaining an extra dollar is more likely than losing a dollar; this is known as the probability premium. Absolute risk aversion is suitable for the comparison of options whose outcomes are absolute changes in wealth, and relative risk aversion where outcomes are percentage changes in wealth. The degree of aversion could be constant, increasing, or decreasing with respect to wealth. A consumer or investor generally takes more risks as they became wealthier, and their preferences can be reasonably approximated by *decreasing absolute risk aversion*.

Matte et al. ~~(2017)~~(2017) assumes that on average a public agency water manager is more likely to exhibit *constant absolute risk aversion* (CARA). For example, we assume that ~~their~~the managers preference for precise forecasts (risk aversion) remains fixed even if the possible losses from one decision are ~~much~~ larger than another decision. In this case a utility function satisfying these properties can be defined by

~~$\mu(E) = -\frac{1}{A}\exp(-A \cdot E)$~~ $\mu(E;A) = -\frac{1}{A}\exp(-A \cdot E)$ (5)

~~where A is the Arrow-Pratt coefficient of absolute risk aversion and E is the economic outcome (Mas-Colell, 1995). Babcock et al. (1993) cautions against interpreting the risk aversion coefficient directly and notes the importance of considering how perception of risk aversion depends on the possible loss. A more interpretable measure which allows comparison between studies with different losses is the *risk premium*; the proportion of loss a decision-maker would pay to eliminate a decision and replace it with a certain outcome. The method introduced here can use any utility function, such as *constant relative risk aversion* which was used by Katz and Lazo (2011).~~

~~The economic model used in this study is a simplified version of that used by Matte et al. (2017) which determines the net outcome from a cost-loss decision to allow systematic comparison with REV, however any economic model can be used. The model used by Matte et al. (2017) can consider intangible damages, distributing spending over multiple lead-times, and calibration to monetary units, and damages informed by flood studies. We are primarily interested in a relative measure of~~

Commented [RL22]: 1.16 (and all other equations with m as a subscript)

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Commented [RL23]: 1.17

Formatted: Font: +Headings (Times New Roman)

Field Code Changed

~~forecast value which can be used more generally for different decision-makers and locations rather than the absolute monetary value of a specific decisions.~~

where the parameter A is the Arrow-Pratt coefficient of absolute risk aversion and E is the economic outcome (Mas-Colell, 1995). Babcock et al. (1993) cautions against interpreting the risk aversion coefficient directly and notes the importance of considering how perception of risk aversion depends on the possible loss. A more interpretable measure which allows comparison between studies with different losses is the *risk premium*; the proportion of loss a user would pay to eliminate a decision and replace it with a certain outcome (Pratt, 1964). The method introduced here can use any utility function, such as *constant relative risk aversion,* which was used by Katz and Lazo (2011).

The economic model used in this study is a simplified version of that used by Matte et al. (2017) which determines the net outcome from a cost-loss decision. The Matte et al. (2017) method considers calibration to monetary units, damages informed by flood studies, intangible damages, and distributed spending over multiple lead-times. Our method is less concerned with the absolute monetary value of forecasts for a specific decision, and instead focuses on the relative value of one forecast over an alternative. This leads to a metric which is more generally applicable and comparable across different users, decisions, forecasts methods, and forecast locations. A cost-loss economic model is required to compare results with REV, and is used in this study; however the RUV method is flexible in that any economic model could be used.

For a state of the world ~~m~~ m at a specific timestep ~~t~~ t, with damages ~~$d(m)$~~ $d_t(m)$, cost to mitigate the damages ~~C_t~~ C_t, and amount of damages avoided ~~$b_t(m)$~~ $b_t(m)$, the outcome is given by

~~$E_{t,m} = b_t(m) - d(m) - C_t$~~ $$E_t(m) = b_t(m) - d_t(m) - C_t \quad (6)$$

The benefit function ~~$b(m)$~~ $b_t(m)$ specifies the damages avoided from taking action to mitigate them,

~~$b_t(m) = \min(\beta \cdot C_t, d(m))$~~ $$b_t(m) = \min\left(\beta \cdot C_t, d_t(m)\right) \quad (7)$$

where the spending leverage parameter ~~β~~ β controls the extra damages avoided for each dollar spent. This is a similar concept (albeit inverted) to the cost-loss ratio ~~α~~ α in the REV metric. The damage function ~~$d(m)$~~ $d_t(m)$ relates the ~~streamflow magnitude~~ states of the world to the economic damages and must be specified for the decision of interest. This economic model assumes that benefits increase linearly as more is spent on damage mitigation, followed by a loss if the spend amount is greater than the damages.

The optimal amount ~~C_t^{π}~~ $\overline{C}_t$ to spend ~~at~~ timestep ~~t~~ t can be found by maximising the expected utility following substitution of Eq. (5)-(7) into Eq. (4),

Field Code Changed
Field Code Changed
Commented [RL24]: 1.18
Commented [RL25]: 2.12
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Commented [RL26]: 1.19 (and every other equation where notation improved)
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Commented [RL27]: 1.20
Field Code Changed
Field Code Changed

$$C_t^{\pi} = \underset{C_t}{\operatorname{argmax}}\ U\left(\tilde{E}_t\right)$$
$$= \underset{C_t}{\operatorname{argmax}} \sum_{m=1}^{M} -\frac{p_t^m}{A} \exp\left[-A \cdot \left(\min\left(\beta \cdot C_t, d(m)\right) - d(m) - C_t\right)\right]$$
$$\bar{C}_t = \underset{C_t}{\operatorname{argmax}}\ U_t\left(E_t\right)$$
$$= \underset{C_t}{\operatorname{argmax}} \sum_{m=1}^{M} -\frac{p_{t,m}}{A} \exp\left[-A \cdot \left(\min\left(\beta \cdot C_t, d_t(m)\right) - d_t(m) - C_t\right)\right] \quad (8)$$

This optimal spend amount for each timestep must be found ex ante, that is before the event has taken place, when the future state of the world is unknown, but a forecast is available. The probabilistic forecast (for some lead-time) is used to determine the forecast likelihood of each state occurring and calculate the ex ante expected utility ~~$U\left(\tilde{E}_t\right)$~~ $U_t\left(E_t\right)$ in Eq. (8). The optimal amount to spend on mitigation is the amount which leads to the largest ex ante expected utility.

The utility can also be calculated ex post, after the event has taken place, and a singular state of the world is known (streamflow observation). This leads to the following expression for the ex post utility after substitutions into Eq. (4)

~~$\Upsilon\left(E_t\right) = \mu\left(\min\left(\beta \cdot C_t^{\pi}, d\left(m_t^o\right)\right) - d\left(m_t^o\right) - C_t^{\pi}\right)$~~

$$\Upsilon\left(E_t\right) = \mu\left(\min\left(\beta \cdot \bar{C}_t, d_t\left(\bar{m}_t\right)\right) - d_t\left(\bar{m}_t\right) - \bar{C}_t\right) \quad (9)$$

where ~~$\Upsilon\left(E_t\right)$~~ $\Upsilon\left(E_t\right)$ is the ex post utility, ~~C_t^{π}~~ $\bar{C}_t$ is the spend amount that was found ex ante, ~~m_t^o~~ $\bar{m}_t$ is the state of the world associated with the observed flow at timestep ~~t~~ t. The ex post utility quantifies the benefit a ~~decision-maker~~user would have gained if they spent ~~C_t^{π}~~ $\bar{C}_t$ on mitigating the damages which occurred as a result of the observed flow. It's important to note that since utility is an ordinal quantity that represents a ~~decision-maker's~~user's preference over the possible decision outcomes, the utilities can be compared but the actual value is noninterpretable. The ex post utility is used in the RUV metric introduced in Sect. 3.

Three ex post metrics were used in Matte et al. ~~(2017)~~(2017) to quantify forecast value using spend amounts found ex ante. They use economic variables (utility, avoided losses and amount spent) averaged over forecasts spanning an historical period. None of these metrics are equivalent or directly comparable to REV and their results were not parameterised by an equivalent of the cost-loss ratio. The mathematical form and interpretation of these 3 metrics are included in the Supplement.

Expected Utility Theory can be used to model more decisions with more realism than is possible with the strong assumptions of REV. However, the economically relevant metrics and parameterisation used to quantify forecast value by Matte et al. ~~(2017)~~(2017) pose a challenge when comparing the outcomes from the two methods.

Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed

~~3~~ 3 Relative Utility Value

This section introduces a new metric which allows direct comparison of the results quantified by the two alternative forecast value approaches described in Sect. 2~~.~~. It aligns the two approaches and allows comparison using the Value Diagram, which is familiar to the environmental modelling verification community and a compelling communication tool. RUV is inspired by REV and skill scores, but with terms based on the ex post expected utility.

~~RUV =~~ $$\mathrm{RUV} = \frac{\mathop{\mathbb{E}}_{t\in T}\left[\Upsilon\left(E_t^{r}\right)\right] - \mathop{\mathbb{E}}_{t\in T}\left[\Upsilon\left(E_t^{f}\right)\right]}{\mathop{\mathbb{E}}_{t\in T}\left[\Upsilon\left(E_t^{r}\right)\right] - \mathop{\mathbb{E}}_{t\in T}\left[\Upsilon\left(E_t^{p}\right)\right]} \quad (10)$$

where ~~$\mathop{\mathbb{E}}_{t\in T}\left[\Upsilon\left(E_t^{\bullet}\right)\right]$~~ $\mathop{\mathbb{E}}_{t\in T}\left[\Upsilon\left(E_t^{\bullet}\right)\right]$ is the expected value of the ex post expected utility from Eq. (9) over a set of observations and either forecast (~~f~~ f), reference ~~climatology~~baseline (~~r~~ r), or perfect information (~~p~~ p). A nice feature of RUV is that it uses the whole probabilistic forecast and does not first convert it to a deterministic forecast like REV.

RUV has all the benefits and familiarity of REV but is a more flexible way to quantify forecast value. Any economic model or form of risk aversion can be used to construct the expected utility terms required by RUV because it is built on the Expected Utility Theory framework. In this paper we focus on the method with the economic model detailed by Eq. (6) and (7), and risk aversion in Eq. (5). If RUV is parameterised using ~~$\beta = \frac{1}{\alpha}$~~ $\beta = \frac{1}{\alpha}$ and visualised on a Value Diagram it can be interpreted in the same way as an REV curve. The flexibility of the utility framework allows the user to make explicit choices about suitable approximations to model the decision problem. This can be accomplished by modifying the economic model, damage function and risk aversion through Eq. (5), (6) and (7) when used to calculate RUV. These assumptions can then be evaluated and extended with additional information if available. Unlike REV using Eq. (2), additional evaluation information is available for each timestep such as the amount spent, damage avoided and economic utility. This may benefit a user applying alternative economic models and tuning damage functions to match real-world data, as they would require the amount spent and damages incurred at individual time steps to determine the components are behaving as expected. Additionally, a user who has finite funds to spend on mitigation and wants to determine when their budget will be exhausted would require investigation of spend and damage amounts at individual time-steps.

3.1 Relationship between RUV and REV

Figure 2 contrasts the processes used by REV and RUV to quantify the value of probabilistic forecasts. Note that RUV uses the same inputs as REV and leads to the same output, however RUV allows the economic model, damage function and risk aversion to be explicitly specified. The internal process is very similar except RUV maximises utility rather than minimises expense.

Formatted: Bullets and Numbering

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Commented [RL28]: 2.13

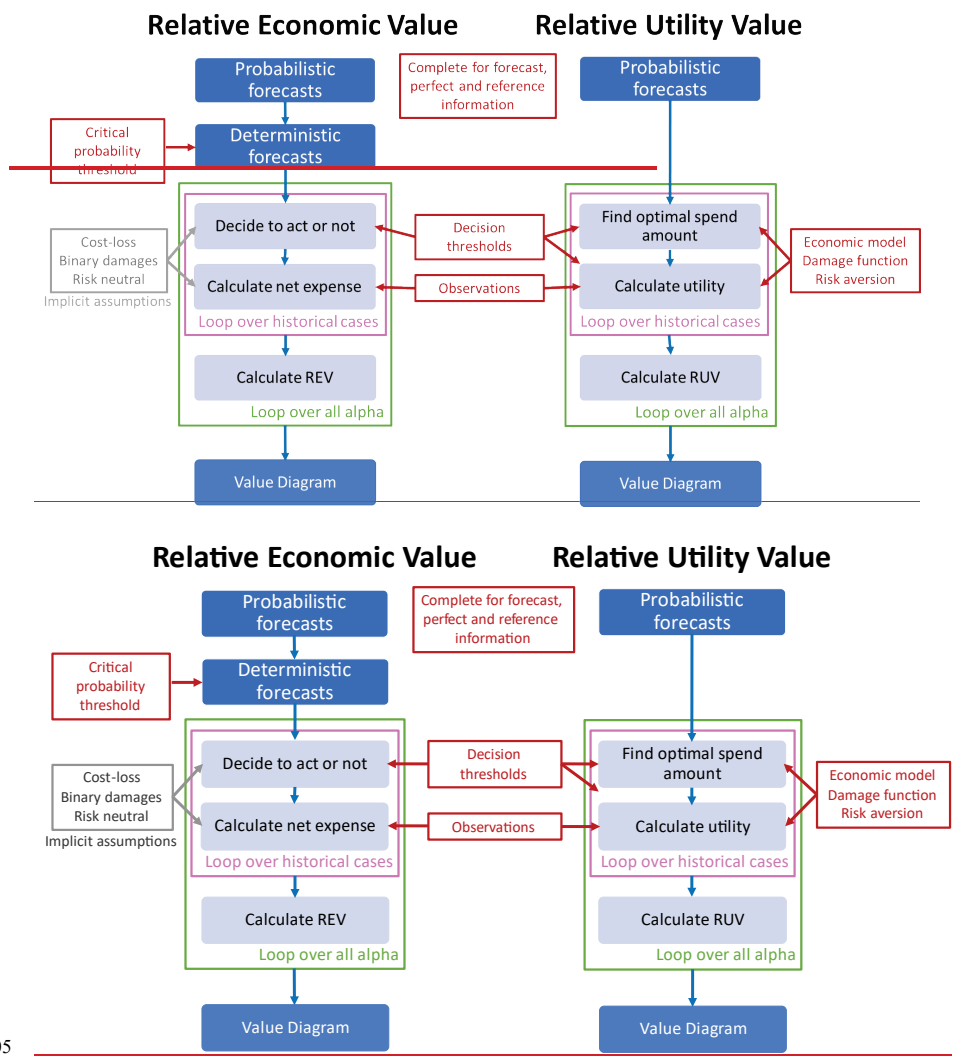


Figure 2: Flowcharts showing the process followed to quantify the value of probabilistic forecasts using either RUV with an optimisation approach to decision making, or REV using the threshold-approach with a specific critical probability threshold. The

sub-processes in the pink boxes are repeated for forecast, perfect, and reference information before being used to calculate REV and RUV. In practice, REV is calculated using Eq. (2~~) is used to calculate REV)~~, which is based on a contingency table with an assumption that it has converged to the long-run performance of the system.

Unlike REV, there is no analytical solution for RUV due to the optimisation step in Eq. (8) unless assumptions are placed on the decision context. When the following 5 assumptions are applied to RUV it is equivalent to REV.

1. Binary damage function is used which is a positive value for the losses above the decision threshold, and 0 otherwise,
2. ~~Decision-makers~~Users are risk neutral as specified by a linear utility function,
3. Forecasts are deterministic with the probability of flow above the threshold always either 1 or 0,
4. The historical frequency of the binary event is used as the reference baseline,
5. All possible losses are avoided.

The mathematical justification for these assumptions and a proof of the equivalence is detailed in Appendix A and the Supplement. Note that when applying these assumptions, the core RUV method illustrated in Figure 2 remains the same but the probabilistic forecast is first converted to a deterministic forecast. Table 2 summarises how decision concepts are represented in each forecast value method and demonstrates the enhanced flexibility of the RUV metric.

Table 2: Comparison of REV and RUV Forecast value methods for defining decisions and ~~decision-maker~~user characteristics

	Relative Economic Value (REV)	Relative Utility Value (RUV)
Level of damages	Fixed loss (dimensionless) Equivalent to step damage function	Damage function is flexible and can be tailored to decision
~~Mitigation~~Level of spending (expense) required to mitigate damages	Fixed cost (dimensionless) Equivalent to fixed spend amount	Spend amount is optimised and varies at each timestep
~~Exposure to damages~~Relative expense of mitigation	Cost-loss ratio	Spending-leverage parameter
Aversion to risk	Always risk neutral	Level of risk aversion and type of utility function can vary
Decision types	Binary	Binary, multi-categorical or continuous-value
Forecast value baseline	Historical event frequency	Any alternative forecast

Formatted Table

Probabilistic decision-making	Threshold-approach	Optimisation-approach or threshold-approach
Economic model	Fixed cost-loss	Economic model is flexible and can be tailored to decision
Interpretation	Value Diagram	Value Diagram

Formatted Table

~~4 Methodology~~

4 ~~A~~Illustrative case study

An illustrative case study is used to ~~determine how~~demonstrate the application of RUV for quantifying the value of ~~probabilistic~~sub-seasonal streamflow forecasts ~~change~~. A series of experiments is used to explore the sensitivity of forecast value to some aspects of decision context, specifically the decision types, users with different ~~decision types, decision makers~~relative expense of mitigation and different levels of risk aversion, and decision-making approaches. A targeted approach is adopted to contrast the RUV and REV methods and illustrate the impact of decision characteristics, rather than an exhaustive evaluation of the value of the specific forecasts used.

Commented [RL29]: 2.15

~~4.1 Background~~

~~Water resource management and the equitable distribution of water to competing stakeholders is challenging due to long-term decreasing trends in available surface water (Zhang et al., 2016), increasing high intensity storm events (Tabari, 2020), river basins overallocated to irrigated agriculture (Grafton and Wheeler, 2018), and deteriorated river system dependant ecosystems (Cantonati et al., 2020). Such challenges to decision-making may be assisted by subseasonal streamflow forecasts and quantifying forecast value would help adoption. For forecasts to be useful they need lead-times long enough to account for river travel-times and decision overheads. Subseasonal forecasts with lead-times out to 30 days would assist management of storages where a release leads to an impact far downstream. For example, agencies operating in the southern Murray-Darling Basin of Australia, such as the Murray-Darling Basin Authority (MDBA) and Goulburn-Murray Water (GMW), make such decisions and may benefit from streamflow forecasts for the Enhanced Environmental Water Delivery method (Murray-Darling Basin Authority, 2017). When operational decisions are informed with probabilistic forecasts the threshold-approach is used with a set of fixed critical probability thresholds, and a degree of risk aversion is implicitly assumed (personal correspondence with MDBA). As far as the authors are aware, the relative value of streamflow forecasts for these decisions and decision-maker characteristics has not been previously quantified.~~

~~4.2 Location~~

~~4.1 Results are presented for the water level station Biggara (401012) on the Murray River in the southern Murray-Darling Basin, Australia. Biggara is upstream~~ Study region and catchment

Our case study explores the value of subseasonal streamflow forecasts at the water level station Biggara (401012) on the Murray River in the southern Murray-Darling Basin, Australia.

Agencies operating in the southern Murray-Darling Basin of Australia, such as the Murray-Darling Basin Authority (MDBA) and Goulburn-Murray Water (GMW), make releases from storages, which have impacts far downstream. Storage management decisions may benefit from subseasonal forecasts, with lead-times out to 30 days, and assist Enhanced Environmental Water Delivery (Murray–Darling Basin Authority, 2017). Currently, when operational decisions are informed with probabilistic forecasts the threshold-approach is used with a set of fixed critical probability thresholds, and a degree of risk aversion is implicitly assumed (personal correspondence with MDBA). As far as the authors are aware, the relative value of streamflow forecasts for these decisions and user characteristics has not been previously quantified.

The Biggara station has particular significance for water resource management in this region as it is located upstream of Hume Dam, a major reservoir used for environmental water releases, irrigated agriculture, and town supply. It is in a temperate region, has a contributing area of 1,257 km^2, a mean rainfall of 1,158 mm/year, and mean runoff of 361 mm/year.

Commented [RL30]: 2.16

Commented [RL31]: 1.22

~~4.3~~4.2 Streamflow forecasts

Formatted: Bullets and Numbering

~~Daily streamflow forecasts are generated using the following method, which demonstrated good performance at subseasonal time horizons in earlier studies (McInerney et al., 2020). We generated 30-day ensemble forecast time series (100 members) starting on the 1st of each month over the period 1991 to 2012. Raw streamflow forecasts were simulated using the GR4J rainfall-runoff model forced by rainfall from the Australian Community Climate and Earth-System Simulator Seasonal (ACCESS-S1) which had been post-processed using the Rainfall Post-Processing for Seasonal forecasts method (RPP-S) (Perrin et al., 2003; Hudson et. al., 2017; Schepen et al., 2018). Final streamflow forecasts were generated by post-processing the raw forecasts using the Multi-Temporal Hydrological Residual Error (MuTHRE) model (McInerney et al., 2020). Post-processing ensured that the statistical properties of the forecasts closely match the observations for all lead-times and accumulations, leading to forecasts which are sharp, reliable, and unbiased. Forecasts with these characteristics can be described as seamless in the sense that they perform well at different time horizons and time resolutions. Further information on these forecasts can be found in McInerney et al. (2020).~~

Daily streamflow forecasts are generated using the following method which demonstrated good performance at subseasonal time horizons in earlier studies (McInerney et al., 2020, 2022). We generated 30-day ensemble forecast time series (100 members) starting on the 1st of each month over the period 1991 to 2012. Raw streamflow forecasts were simulated using the GR4J rainfall-runoff model (Perrin et al., 2003), forced by rainfall from the Australian Community Climate and Earth-System Simulator Seasonal (Hudson et. al., 2017) that had been pre-processed using the Rainfall Post-Processing for Seasonal forecasts

method (Schepen et al., 2018), and potential evapotranspiration from the Australian Water Availability Project (Jones et al., 2009). Final streamflow forecasts were generated by post-processing the raw streamflow forecasts using the Multi-Temporal Hydrological Residual Error (MuTHRE) model (McInerney et al., 2020). Post-processing ensured that the statistical properties of the streamflow forecasts closely match the streamflow observations. The MuTHRE model was chosen for post-processing because it provides "seamless" forecasts that are (statistically) reliable and sharp across multiple lead-times (0-30 days) and aggregation time scales (daily to monthly). Further information on the forecasts used in this study can be found in McInerney et al. (2020), and further method improvements to enhance seamless performance in McInerney et al. (2021).

Commented [RL32]: 2.17

Commented [RL33]: 1.25

Commented [RL34]: 1.24

Commented [RL35]: 2.18

~~4.4~~4.3 Decision types

Formatted: Bullets and Numbering

~~A binary decision of flow exceeding a single threshold can be considered the simplest for a decision-maker to manage, flow exceeding the height of a levee for example. A multi-categorical decision with more than two classes introduces additional complexity for the decision-maker to consider. An example of this is a minor, moderate, and major flood classification which correspond to increasing categories of impact. A mitigation decision based on continuous flow is the limiting case of a very large number of flow classes. An example is adjusting dam releases to match storage inflow during flood operations. A binary decision has traditionally been used as the prototypical model of decision making in decision-theoretic literature and is a limiting assumption of REV (Katz and Murphy, 1997). Decisions involving more flow classes are more complex for decision-makers to reason about as there are more possible outcomes to consider, but they are an essential feature of many real-world decisions and cannot be ignored.~~

~~Three types of decisions have been included in the~~Decisions involving more than two flow classes are an essential feature of many real-world decisions (see examples in Sect. 1). Three types of decisions are considered in the illustrative case study: (i) binary decisions with flow above a single threshold, either the top 25% of top 10% of the observation record; (ii) multi-categorical decisions with flow in 5 classes over a range of thresholds; and (iii) continuous-flow decisions using flow from whole flow regime. These thresholds are indicative of decisions ~~which~~that depend on moderate to high flow at Biggara, such as operational airspace management of the Hume Dam or minor inundation upstream of Yarrawonga Weir when coinciding with a dam release.

Commented [RL36]: 2.19

~~4.5~~4.4 Economic damages

Formatted: Bullets and Numbering

The relationship between damages and flow in Eq. (6) and (7) when applying the RUV metric is specified using a non-dimensional logistic function,

$$\sout{d(q) = \frac{\delta}{1+\exp(-k(q-\tau))}}\ d(q;\delta,k,\phi) = \frac{\delta}{1+\exp(-k(q-\phi))} \qquad (11)$$

Field Code Changed

The logistic function can be parameterised to have very similar behaviour to the Gompertz curve used in flood damage studies and used by Matte et al. ~~(2017), with $d(q)$ representing the cumulative damages incurred from all flow up to q (Li et al.,~~

~~2016a). It was parameterised to reasonably characterise losses from high flow events; no damages when flow is zero, increasing quickly from around the top 20% of flow, and approaching 1 at very high values above the top 1% of flow. The following parameter set was found to be suitable; $\delta = 1$, $k = 1$ and τ equal to the value corresponding to the top 1% of observed historical flow, see Figure 3 and~~ (2017), with $d(q)$ representing the cumulative damages incurred from all flow up to q (Li et al., 2016a). It was parameterised to reasonably characterise losses from high flow events; no damages when flow is zero, increasing quickly from around the top 20% of flow, and approaching 1 at very high values above the top 1% of flow (see Figure 3a). These assumptions were reproduced with the following parameter set; , $\delta = 1$, $k = 0.07$ and ϕ equal to the value corresponding to the top 1% of observed historical flow.~~Sect. 6.3.~~

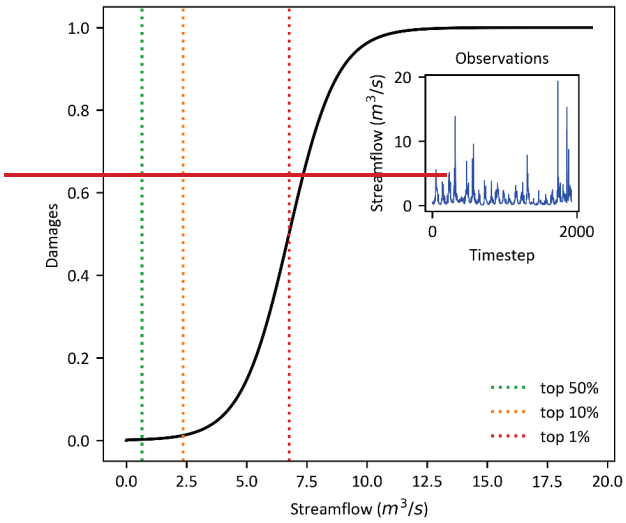
Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed



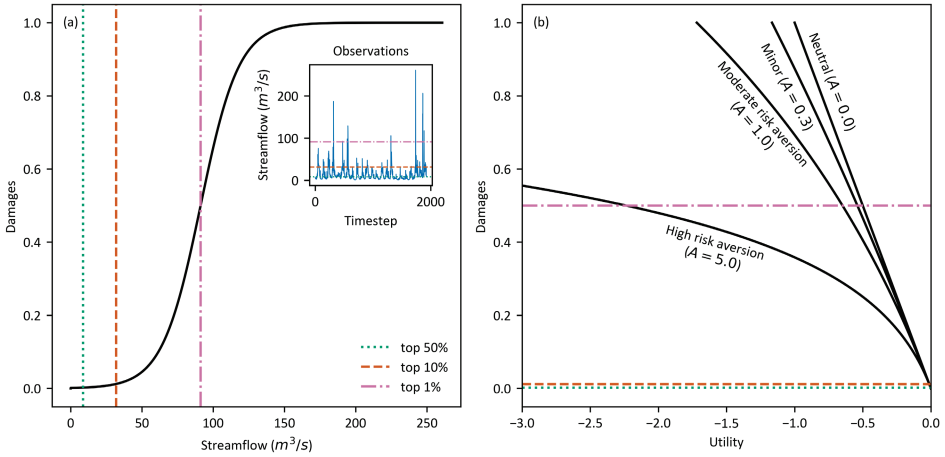


Figure 3: ~~Damage~~(a) Example damage function used in the illustrative case study based on a logistic curve with an inflection point at the top 1% of observed flow., and (b) corresponding CARA utility function with 4 levels of risk aversion (limited to $-3 \leq$ utility ≤ 0 and aligned at zero utility for visual clarity).

Field Code Changed

~~4.6~~4.5 Risk aversion

It is difficult to precisely know a ~~decision-maker's~~user's level of risk without a history of prior decisions. Moreover, it would be incorrect to assume that all ~~decision-maker's~~ users share the same level of risk. Therefore, a range of risk aversions have been considered to illustrate its impact on forecast value. In this study we have used risk aversion coefficients ~~$A \in \{0, 0.3, 1, 5\}$~~ $A \in \{0, 0.3, 1, 5\}$, which correspond to risk premiums of ~~$\theta \approx \{0\%, 15\%, 44\%, 86\%\}$~~ $\theta \approx \{0\%, 15\%, 43\%, 86\%\}$ for a CARA utility function with maximum losses of ~~$\delta = 1$~~ $\delta = 1$ ~~(Babcock et al., 1993). These~~ (Babcock et al., 1993), see. Figure 3b shows that the curvature of $\mu(\cdot)$ increases with increasing risk aversion, and this leads to an increasingly rapid decline in utility from damages. The 4 risk aversion coefficients represent ~~decision-makers~~users who are neutral, ~~slightly~~minorly, moderately, and highly risk averse~~.~~ respectively. When risk premiums are considered, our range of risk aversion coefficients is similar to those used by Tena and Gómez ~~(2008)~~(2008) and Matte et al. ~~(2017).~~(2017). Finding appropriate values of risk aversion for a specific ~~decision-maker~~user is beyond the scope of this study, but would be highly beneficial in user-focused forecast value studies.

Formatted: Bullets and Numbering

Commented [RL37]: 1.27

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Commented [RL38]: 1.17

Formatted: Font: Bold

4.74.6 Experiments

The value of the subseasonal forecasts are quantified using the RUV and REV metrics. Experiments are performed over the dimensions of forecast lead-time, decision type, decision making approach, metric, and ~~decision maker~~user risk aversion. Streamflow forecasts from multiple daily lead-times were grouped together to quantify forecast value over 7-day and 14-day forecast horizons. Grouping lead-times together simplifies the introduction of RUV and comparison of its salient features with REV; however, for practical applications there may be benefits for evaluating forecast value at specific lead-times of interest. A fixed climatology based on all observed values in the record is used for the reference baseline of RUV to align with that used in REV. Table 3 summarises the specific attributes used for each figure, with the key dimension highlighted as red text.

Table 3: Dimensions of forecast value problem used for each figure. Key dimension introduced in each figure is highlighted with red text.

Experiment purpose	Lead-times (days)	Decision type	Decision thresholds	Decision-making approach	Metric	Risk aversion
Experiment 1: Equivalence of REV and RUV, and impact of fixed probability thresholds. Moderate flow example. (Figure 4)	1-7	Binary	Top 25%	*Threshold*	*REV* *RUV*	0
Experiment 2: Contrast decision-making approaches. Moderate flow example. (Figure 5)	1-7	Binary	Top 25%	*Threshold* *Optimisation*	RUV	0
Experiment 3: Subseasonal forecast value for different decision types. High flow examples. (Figure 6)	1-7 8-14 15-30	*Binary* *Multi-categorical* *Continuous-flow*	Top 10% ~~Top 20%, 15%, 10%, 5%~~ All flow	Optimisation	RUV	0
Experiment 4: Impact of risk aversion ~~of~~ on forecast value. High flow examples. (Figure 7)	1-7	Binary Multi-categorical Continuous-flow	Top 10% Top 15%, 10%, 5%, 1% All flow	Optimisation	RUV	*0, 0.3, 1, 5*
Experiment 5: Key driver of impact of risk aversion on forecast value. (Figure 8)	1-7	Binary	~~*All flow*~~*Thresholds from bottom 5% to top 0.03%*	Optimisation	RUV	0, 0.3, 1, 5

Formatted: Bullets and Numbering

Commented [RL39]: 2.4b

Commented [RL40]: 1.28 & 2.21

Formatted Table

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Italic

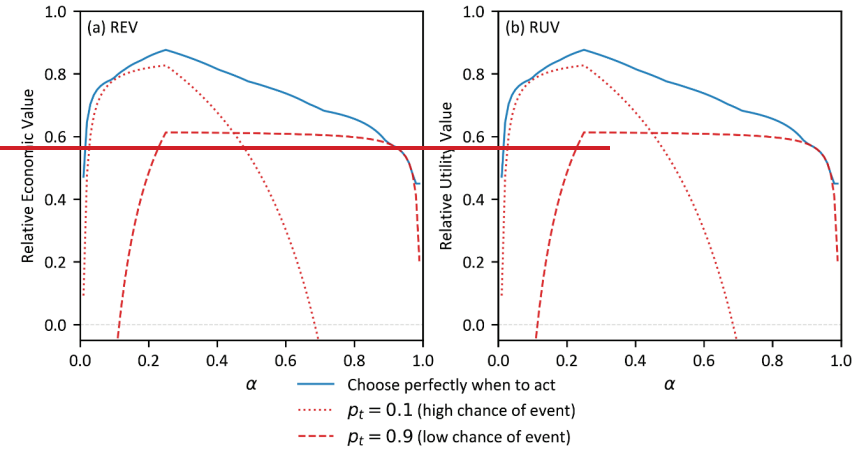
Formatted: Font: Italic

~~5~~ 5 Results

Formatted: Bullets and Numbering

5.1 Experiment 1: Equivalence of RUV and REV, and impact of fixed probability threshold

In ~~experiment~~Experiment 1, forecast value has been quantified using REV and RUV with the assumptions detailed in Sect. 3.1: binary damage function, risk neutral ~~decision maker~~user, deterministic forecasts, event frequency for reference baseline, and all losses avoided. As expected, Figure 4 demonstrates that the results are identical between the two methods.



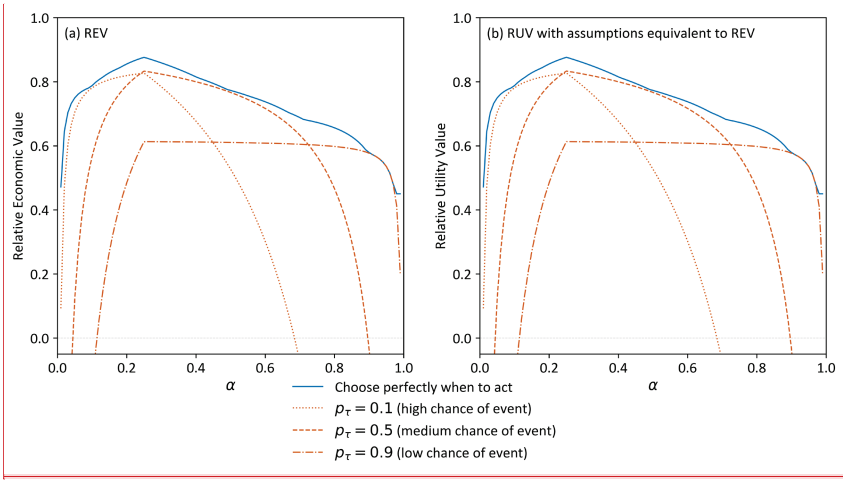


Figure 4: Forecast value quantified using (a) REV and (b) RUV with assumptions enforced, and the threshold approach for decision-making. With a binary decision of flow exceeding the top 25% of observations, subseasonal forecasts from the first week of lead-times, and a risk neutral ~~decision-maker.~~user. Critical probability thresholds for the ~~three~~four curves are the value leading to maximum forecast value, and the 0.1, 0.5, and 0.9 forecast quantile, corresponding to acting when there is a high, medium, or low chance of event occurring respectively.

Commented [RL41]: 1.29 & 2.22

We now explore the detrimental impact on forecast value of using the threshold-approach to convert probabilistic forecasts to deterministic forecasts. Any forecast value method using the threshold-approach needs to select a critical probability threshold ~~p_t~~ p_τ to convert probabilistic forecasts to deterministic forecasts. Figure 4 includes ~~three~~3 curves corresponding to decisions made with different thresholds. The blue line shows the value obtained when the threshold ~~p_t~~ p_τ is chosen to maximise that value at each ~~α~~ α (see Sect. 2.1.2). This is an upper limit that cannot be obtained in practical situations because it implies a ~~decision-maker~~user has either perfect foresight or a perfectly reliable forecast, and ~~$P_t = \alpha$~~ $p_\tau = \alpha$ will lead to maximum value if the forecast is perfectly reliable ~~(Richardson, 2000)~~(Richardson, 2000). The ~~red~~orange lines show how the choice of ~~p_t~~ p_τ can have a dramatic impact on the value of forecasts for a decision, with the dotted line showing forecast value when ~~$p_t = 0.1$~~ ~~and~~ $p_\tau = 0.1$, the dashed line when ~~$p_t = 0.9$~~ $p_\tau = 0.5$, and the dash-dot line when $p_\tau = 0.9$. RUV is negative for some regions of ~~α~~ α, which indicates that those ~~decision-makers~~users should ~~use~~prefer the climatological baseline rather than the forecasts when making decisions.

This result clearly shows that to extract the most value from forecast information a ~~decision-maker needs to consider their exposure to damages α when choosing p_t. For example, when a decision-maker with $\alpha = 0.8$ uses $p_t = 0.9$ they gain significant value from the forecasts (RUV ≈ 0.6), but if they use $p_t = 0.1$ their outcome using forecasts is worse than using the reference climatology (RUV < 0), while for a different decision-maker with $\alpha = 0.1$ the opposite is true. It additionally shows that the Value Diagram used with REV remains a compelling way to visualise how RUV forecast value varies for different decision-makers~~user needs to consider their relative expense of mitigation α when choosing p_τ. For example, when a user with $\alpha = 0.8$ uses $p_\tau = 0.9$ they gain significant value from the forecasts (RUV ≈ 0.6), but if they use $p_\tau = 0.1$ their outcome using forecasts is worse than using the reference baseline (RUV < 0), while for a different user with $\alpha = 0.1$ the opposite is true. This critical dependence of value on p_τ is an established finding for REV (Richardson, 2000; Murphy, 1977) and is not specific to this example; here we illustrate that RUV reproduces it. Figure 4 additionally shows that the Value Diagram used with REV remains a compelling way to visualise how RUV forecast value varies for different users.

This result, and the derivation in Appendix A- and the Supplement, demonstrate that RUV and REV are equivalent when appropriate assumptions are imposed. It demonstrates that REV can be considered a special case of the more general RUV metric.

5.2 Experiment 2: Contrasting the threshold-approach and optimisation-approach for decision making

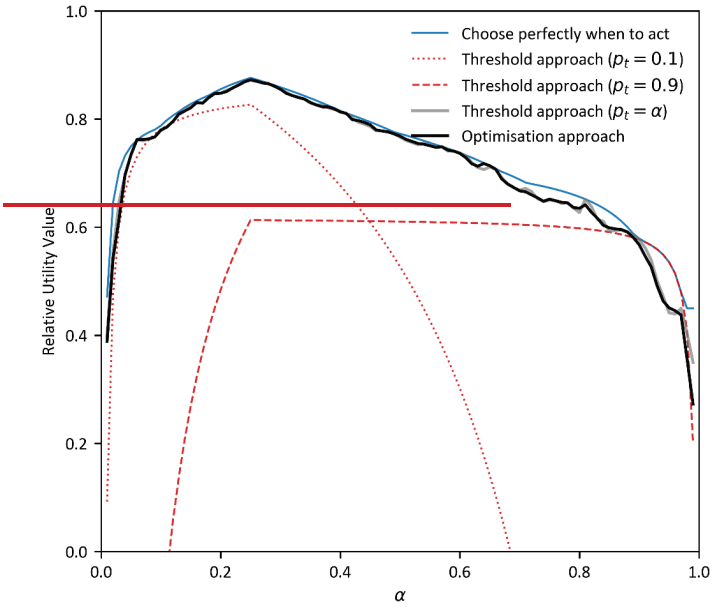
Figure 5 adds two more forecast value curves, generated using RUV, to Figure 4. The black line shows value when the optimisation-approach is used to make spending decisions with the subseasonal forecasts (detailed in Sect. 2.2), and the ~~grey~~pink line shows value when the threshold-approach is used with ~~$P_t = \alpha$~~ $p_\tau = \alpha$. The result demonstrates that making

Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Commented [RL42]: 1.29
Field Code Changed
Field Code Changed
Field Code Changed
Commented [RL43]: 2.3
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Commented [RL44]: 2.4c
Field Code Changed

decisions using either approach provides close to the maximum value possible for all ~~decision-makers~~users (different values of ~~α~~ α). This contrasts dramatically with the threshold-approach using specific fixed values for ~~p_t~~ ~~(red~~ p_τ (orange lines) which only provides maximum value for a very small range of ~~decision-makers~~users.

Field Code Changed

Field Code Changed


Choose perfectly when to act
Threshold approach ($p_t = 0.1$)
Threshold approach ($p_t = 0.9$)
Threshold approach ($p_t = \alpha$)
Optimisation approach
Relative Utility Value
α
0.0
0.2
0.4
0.6
0.8
1.0

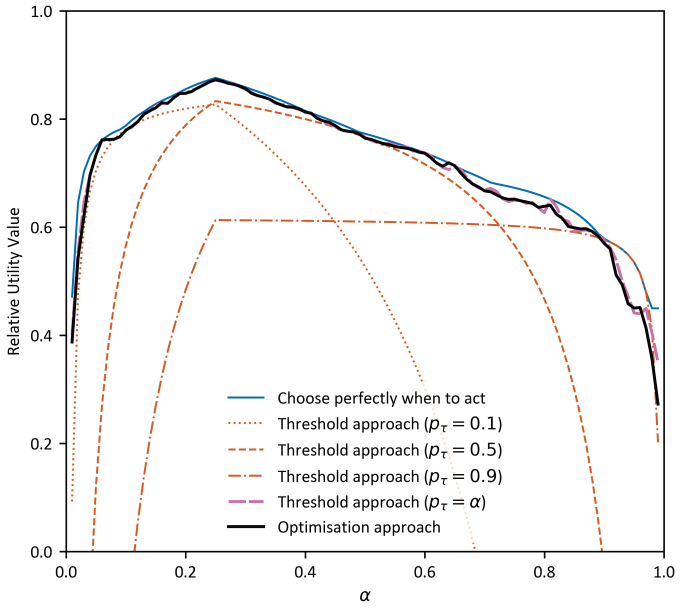


Figure 5: Forecast value quantified using four different approaches to decision-making: the optimisation-approach and the threshold-approach with either perfect critical probability thresholds, specific critical thresholds, or the critical threshold set equal to the ~~alpha~~ α value. A binary decision of flow exceeding the top 25% of observations was used, with subseasonal forecasts from the first week of lead-times, and a risk neutral ~~decision-maker.~~user. Specific critical thresholds are the 0.1, 0.5, and 0.9 forecast quantile, corresponding to acting when there is a high or low chance of the event occurring respectively.

Field Code Changed

Investigations (not shown) indicated that the optimisation and ~~$P_t = \alpha$~~ $p_\tau = \alpha$ curves (black and ~~grey~~pink lines) are non-smooth because of the limited number of events in the observation record, and the small difference between the ~~grey and~~ black and pink lines is due to sampling errors from to the relatively small ensemble ~~sampling error.~~ size. It is notable that forecast value from these two different decision-making approaches are essentially equivalent as illustrated by the closeness of the black and ~~grey~~pink lines in Figure 5. Additional analysis (not shown) found this equivalence to be robust to the type of decisions (binary, multi-categorical, or continuous-flow) ~~and changes in forecast reliability.~~ but not equivalent for risk averse users.

Field Code Changed

Commented [RL45]: 2.24

5.3 Experiment 3: Comparing Forecast value for different types of decisions

Figure 6 presents results for binary~~,~~ (blue-lines), multi-categorical (orange lines) and continuous-flow decisions (green lines) with forecast lead times in separate panels. RUV was calculated for the daily subseasonal forecasts with lead-times pooled

from the 1st week ~~(blue lines~~(Figure 6a), 2nd week ~~(orange lines~~(Figure 6b) and 3rd and 4th weeks combined ~~(green lines).~~(Figure 6c). The ~~decision-maker~~user is assumed to be risk neutral, and the optimisation-approach was used. Overall, the forecasts provide excellent value for these three different decision types over all time-horizons (max 30 days), implying that any ~~decision-maker~~user would likely benefit from using the forecast information over the ~~climatology~~reference baseline. Peak RUV ~~was~~is over 0.8 in the first week for all decision types, and close to 0.7, 0.6, and 0.5 in subsequent weeks for binary, multi-categorical, and continuous-flow decision types respectively. ~~The exception is for decision-makers with high exposure to damages in the 3rd and 4th weeks, where RUV drops below zero above $\alpha = 0.6$ for binary decisions and $\alpha = 0.9$ for multi-categorical decisions.~~ Regardless of the decision type or lead-time, forecasts provide maximum value for users with α close to the probability of the most damaging flow class occurring. For example, for the binary decision the peak RUV value is located at $\alpha = 0.1$, which corresponds with the event frequency of decision threshold used (top 10% of flow).

Field Code Changed

Field Code Changed

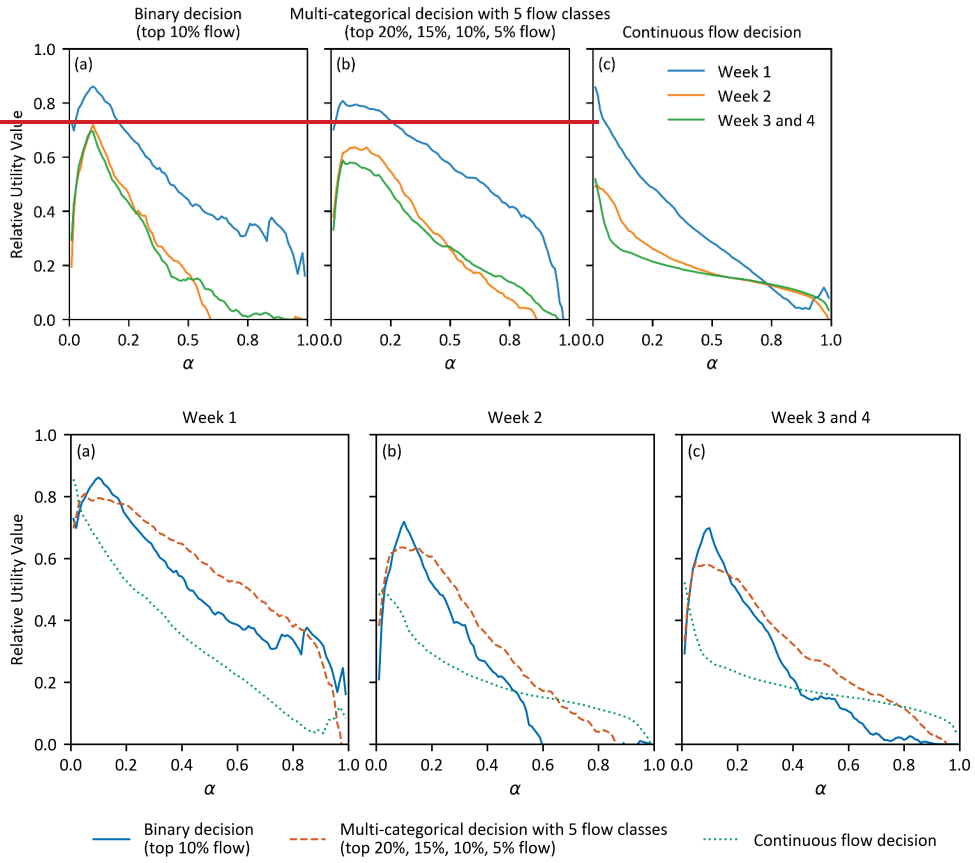


Figure 6: Forecast value for (a) binary decision of flow exceeding the top 10% of observations, (b) flow within 5 classes with thresholds at the top 20%, 15%, 10% and 5% of observations, and (c) continuous-flow. Decisions are made using the optimisation-approach for decision-making with a risk neutral ~~decision-maker~~user, and subseasonal forecasts for the 1st, 2nd and combined 3rd and 4th weeks of lead-times.

~~Almost all decision-makers will experience positive value from incorporating the streamflow forecasts into their decisions across all lead-times and decision types. The only decision-makers who should avoid using the forecasts in all cases are those with very large exposure to damages, a common finding in studies using REV (Roulin, 2007).~~ However,Figure 6 shows that there is important variation in RUV ~~across α, lead-time, and decision type.~~for different decision types. These differences in

Formatted: Superscript

Formatted: Superscript

Formatted: Superscript

Commented [RL46]: 1.30 & 2.36 (and all other figures for the colours/line-styles changes)

Formatted: Superscript

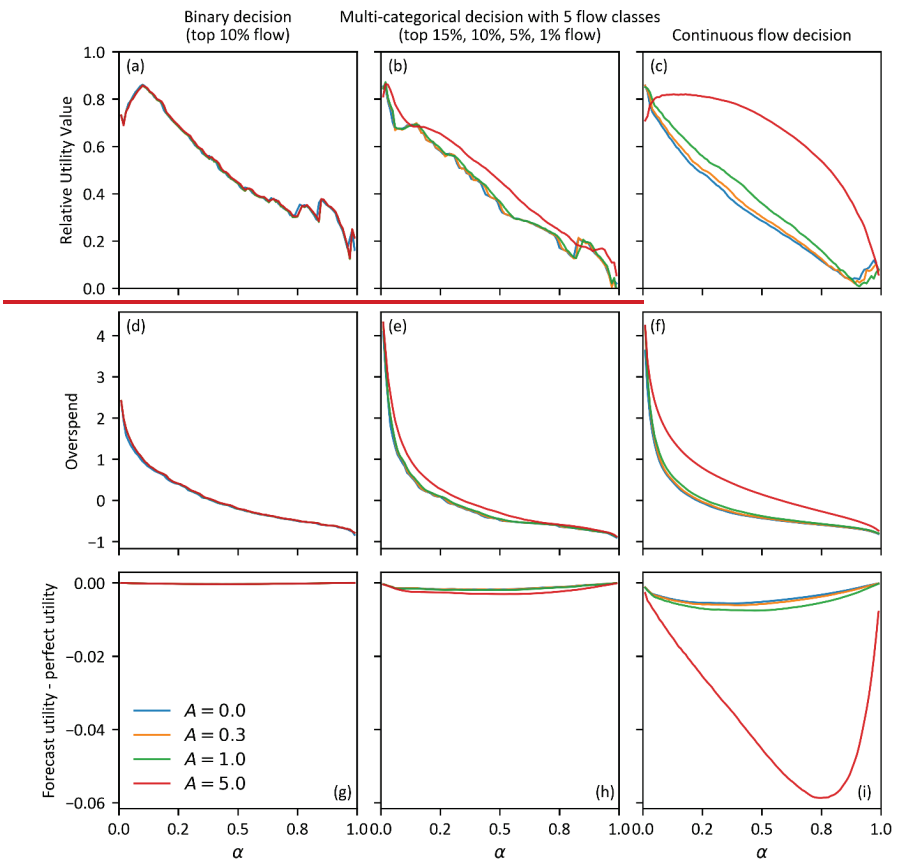
RUV for different decision-types are more pronounced for larger values of α and at longer lead-times. For example, ~~beyond the second~~for users with $\alpha > 0.6$ (lead-time week ~~decision-makers with $\alpha > 0.6$~~2) the RUV is below zero for the binary decision type, but not the multi-categorical or continuous flow decision types. This suggests the users should prefer the reference ~~climatology~~baseline for the binary decision and prefer forecasts for the multi-categorical and continuous-flow decisions. ~~Regardless of the~~This highlights the importance of calculating forecast value using the decision type which matches the decision being assessed.

It is notable that for higher values of α the value of forecasts in weeks 3 and 4 is higher than week 2. While differences are minor, they interestingly appear robust over the multiple decision ~~type or~~ types in this case study. The reduced value of forecasts could possibly be due to lead-time~~, forecasts provide maximum value for decision-makers with α close to the probability of the most damaging flow class occurring. For example, for the binary decision the peak RUV value is located at $\alpha = 0.1$ which corresponds with the event frequency of decision threshold used (top 10% of flow). Forecast~~ dependent differences in forecast reliability and decreasing sharpness of the forecast ensemble at longer lead-times. Another notable feature is that forecast value at small ~~α~~ α is enhanced for continuous-flow decisions relative to the other decision- types. This seems to be because large damages from infrequent extreme events are more adequately mitigated in continuous-flow decisions because a correspondingly large amount is spent when they are forecast correctly.

5.4 Experiment 4: Impact of risk aversion

Experiment 4 contrasts forecast value for a risk neutral ~~decision-maker~~user against ~~3~~three different levels of risk aversion for binary, multi-categorical, and continuous-flow decisions. The results presented in Figure 7 for the RUV metric (first row) as well the overspend (middle row) and utility-difference metrics (last row) used by Matte et al. ~~(2017)~~(2017) which provide insight into the spending decisions and utility respectively. By varying ~~A~~ A in Eq. (5) risk aversion is found to have a ~~moderate~~significant impact on the value of forecasts for ~~the~~highly risk averse users making continuous-flow decisions, a moderate impact for multi-categorical and continuous-flow decisions~~,~~ (except for highly risk averse users), and a minor impact for binary ~~decision types~~decisions (see Figure 7 first row). Increased risk aversion shifts the RUV curve toward users with higher ~~α~~ α, suggesting that risk averse ~~decision-makers~~users with more ~~exposure to damages~~expensive mitigation would benefit more from using forecasts to make their decisions.

Field Code Changed
Field Code Changed
Commented [RL47]: 2.25
Field Code Changed
Field Code Changed
Commented [RL48]: 1.31
Field Code Changed
Commented [RL49]: 2.27
Field Code Changed


Binary decision
(top 10% flow)
Multi-categorical decision with 5 flow classes
(top 15%, 10%, 5%, 1% flow)
Continuous flow decision
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
Relative Utility Value
Overspend
Forecast utility - perfect utility
A = 0.0
A = 0.3
A = 1.0
A = 5.0
α

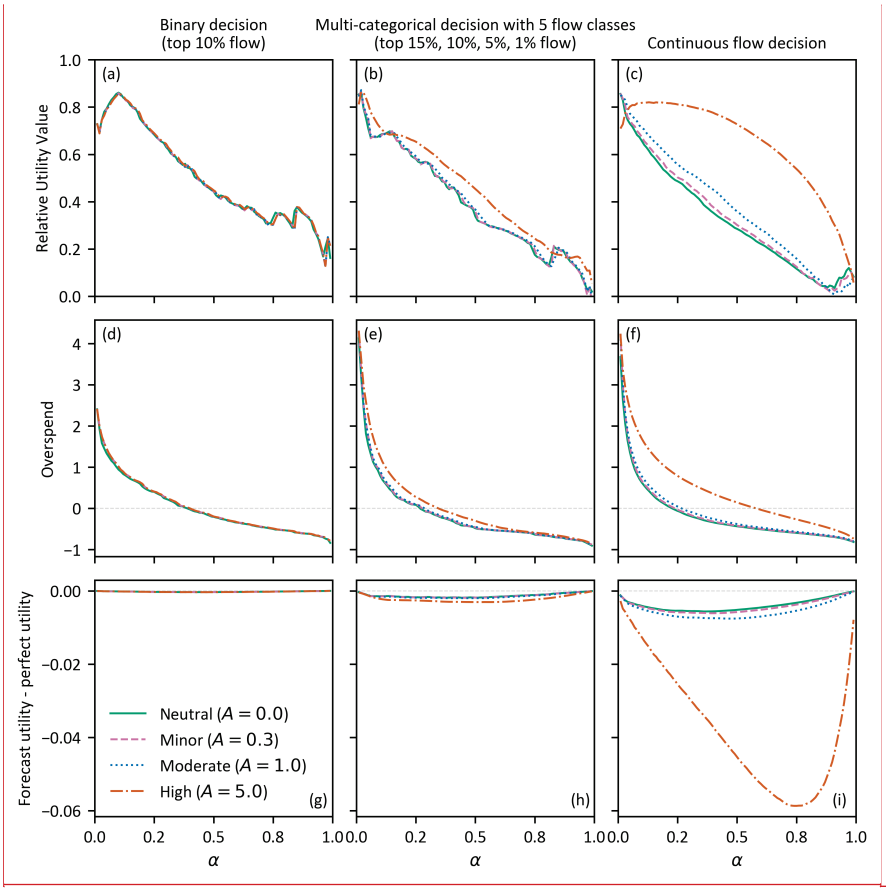


Figure 7: RUV, overspend and utility-difference for different levels of ~~decision-maker~~user risk aversion, for a binary decision of flow exceeding the top 10% of observations (first column), flow within 5 classes with thresholds at the top 15%, 10%, 5%, and 1% of observations (middle column), and continuous-flow (last column). Decisions made using the optimisation approach with subseasonal forecasts from the 1st week of lead-times.

The overspend ~~(~~(Figure 7, middle row) and utility-difference results ~~(~~(Figure 7, last row) indicate that risk aversion has a minor impact on the spending decisions and the resultant utility~~.~~, except for highly risk averse users making continuous-flow decisions. The overspend panels show that regardless of risk aversion, on average a ~~decision-maker~~user will spend more than

Commented [RL50]: 1.33

Formatted: Superscript

Commented [RL51]: 2.27

necessary when their cost of mitigation is small relative to the potential avoided losses (small ~~α).~~ α). Conversely, when ~~α~~ α is large they will underspend on average. When risk aversion is increased, ~~decision makers~~users spend increasingly more. The utility-difference panels (Figure 7, bottom row) show that decisions made using forecasts provide users less utility than decisions made using perfect information, and this decrease in utility increases with risk aversion. As utility is an ordinal measure it is only meaningful to interpret differences within each panel (g), (h), and (i), not between them. This highlights a benefit of the overspend and RUV metrics which are comparable across decision type.

Field Code Changed

Field Code Changed

Commented [RL52]: 1.34

5.5 Experiment 5: Mechanism behind the varying impact of risk aversion

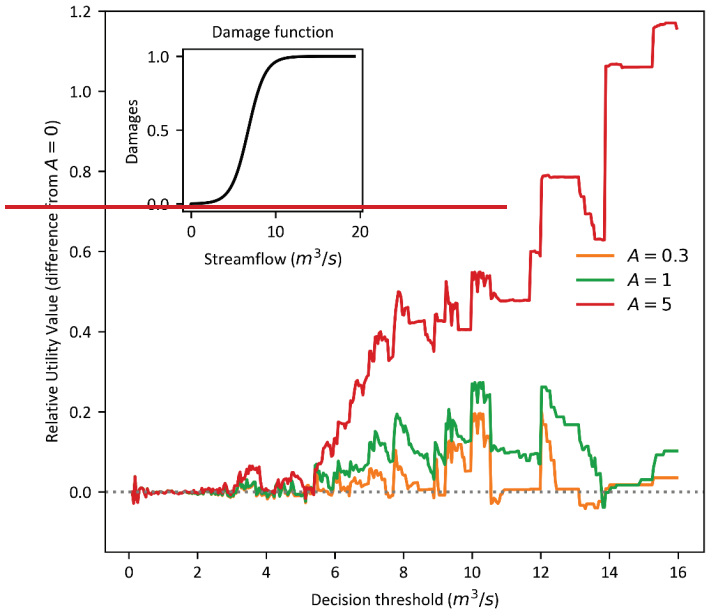
It is notable that the impact of risk aversion in Figure 7 is different for each decision type; minor for the binary decisions, moderate for multi-categorical and continuous-flow, and particularly enhanced for highly risk averse ~~decision makers.~~users. Experiment 5 investigates the mechanism behind this. Figure 8 presents the difference in RUV between risk averse and risk neutral ~~decision makers~~users (y-axis), for a binary decision at a single ~~exposure to damages~~value of ~~$\alpha = 0.2$.~~ α ($\alpha = 0.2$). The binary decision threshold (x-axis) is varied from ~~the 0.1 – 16~~2 – 225 m^3/s (bottom ~~2~~5% to top 0.~~04~~03%) and decisions are made using the optimisation approach with subseasonal forecasts from the 1st week of lead-times. This contrasts with the binary decision in experiment 4 where the decision threshold is fixed at ~~2.4~~32 m^3/s (top 10%) and ~~α .~~ α is varied.

Formatted: Font: Italic

Field Code Changed

Field Code Changed

Formatted: Font: Italic



~~Figure 8: difference in RUV between risk averse ($A > 0$) and risk neutral ($A = 0$) decision-makers (y-axis), for a binary decision at a single exposure to damages of $\alpha = 0.2$. The binary decision threshold (x-axis) is varied from 0.1 – 16 m3/s and decisions are made using the optimisation approach with subseasonal forecasts from the 1st week of lead-times.~~

Below a critical decision threshold of approximately ~~5~~70 m^3/s (top 2% flow) the difference in RUV between any level of risk aversion and risk-neutrality is negligible. Above this value an increasing difference is clear, particularly in the highly risk averse case, with risk averse ~~decision-makers~~users gaining more value from the forecast information than risk neutral. This finding was consistent for multi-categorical decisions of any number of flow classes, all lead-times, and all values of ~~α~~ α except at extreme high and low values (not shown). The specific experimental values (binary decision, $\alpha = 0.2$, 1st week lead-time) were chosen as a representative example and the findings apply for other experimental values. It demonstrates that the decision thresholds used, specifically in relation to the damage function, are the key drivers behind the impact of risk aversion regardless of the decision type. The difference in impact of risk aversion across the different decision types in Figure 7 can therefore be explained by the specific decision thresholds used in relation to this critical value; the binary decision threshold of ~~2.4~~32 m^3/s used in experiment 4 was less than the critical value of ~~5~~70 m^3/s and only a minor impact from risk aversion was found, whereas the top decision threshold for the multi-categorical decision was ~~6.8~~91 m^3/s, above this critical value, and a

Field Code Changed

Field Code Changed

Commented [RL53]: 2.28

moderate impact was found, and an even larger impact was found for the continuous-flow decision which includes contribution from the largest flows.

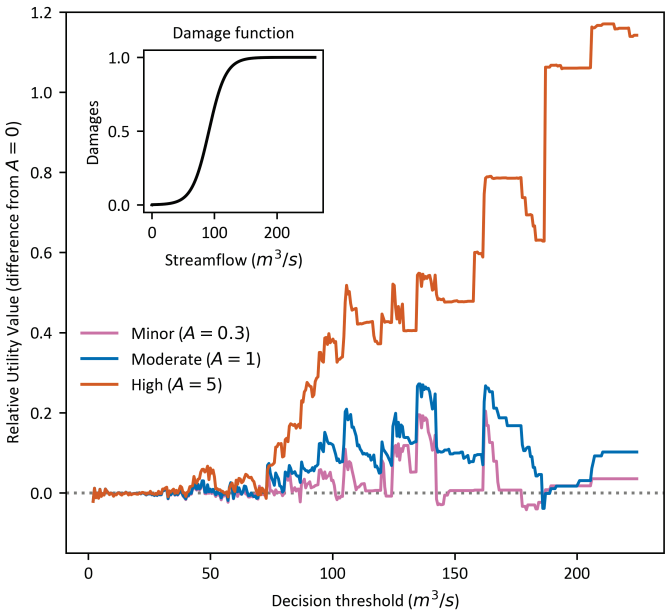


Figure 8: difference in RUV between risk averse ($A > 0$) and risk neutral ($A = 0$) users (y-axis), for a binary decision at a single α value ($\alpha = 0.2$). The binary decision threshold (x-axis) is varied from 2 – 225 m^3/s and decisions are made using the optimisation approach with subseasonal forecasts from the 1st week of lead-times.

Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed

~~6~~ 6 Discussion

Formatted: Bullets and Numbering

~~According to statistical forecast verification metrics, probabilistic streamflow forecasts have been shown to be skilful and statistically reliable (McInerney et al., 2021; Li et al., 2016b). However, their ability to improve decision outcomes has not been extensively established. Additionally, REV, the most frequently used forecast value method, can only be applied to a limited number of real-world decisions. In this paper we develop a new forecast value method, Relative Utility Value (RUV), which is more flexible than REV and can applied to more decision. The flexibility of RUV is demonstrated with a case study using probabilistic subseasonal streamflow forecasts to inform binary, multi-categorical, and continuous flow decisions with~~

~~risk averse decision-makers. The 5 experiments reported in Sect. 5 systematically explore the impact of different aspects of a decision on forecast value: the forecast value method, the probabilistic decision-making approach, types of decisions, decision-maker risk aversion, and the mechanism behind varied risk aversion impact. First, we find that under certain conditions RUV and REV are equivalent, and REV can be considered a special case of the more general RUV method (see Figure 4 and Appendix A). Second, making decisions with fixed critical probability thresholds is leads to maximum forecast value only for a very small set of users, and using an optimisation-based approach makes better use of probabilistic forecast information (see Figure 5). Third, subseasonal forecasts offer more value than a climatological average for almost all lead-times and decision-makers regardless of the decision-type (see Figure 6). And finally, risk aversion has a minor to moderate impact on forecast value (see Figure 7) but the degree of impact is sensitive to the decision context being evaluated. The key mechanism driving this impact is the decision thresholds used relative to the damage function (see Figure 8). This section interprets these results through the lens of forecast users and producers.~~

~~6.1 Benefits of RUV over alternatives~~

~~*Forecast value complements forecast verification.* Unlike forecast verification, forecast value considers the broader context within which decisions are made.~~ Statistical forecast verification metrics have previously been used to show that the probabilistic streamflow forecasts used in this study are reliable and sharp, largely due to the post-processing method employed (McInerney et al., 2021). Other post-processing methods have also demonstrated capability to improve the reliability and sharpness of raw streamflow forecasts (Bogner et al., 2016; Li et al., 2016b; Woldemeskel et al., 2018; Lucatero et al., 2018). However, the ability of these forecasts to improve decision outcomes has not been extensively established. Additionally, REV, the most frequently used forecast value method, can only be applied to a limited number of real-world decisions. In this paper we developed a new forecast value method, Relative Utility Value (RUV), which is more flexible than REV and can be applied to more decisions. The flexibility of RUV is demonstrated with an illustrative case study using probabilistic subseasonal streamflow forecasts to inform binary, multi-categorical, and continuous-flow decisions with risk averse users. The 5 experiments reported in Sect. 5 systematically explore the impact of different aspects of a decision on forecast value: the forecast value method, the probabilistic decision-making approach, types of decisions, user risk aversion, and the mechanism behind varied risk aversion impact. First, we find that under certain conditions RUV and REV are equivalent, and REV can be considered a special case of the more general RUV method (see Figure 4, Appendix A, and the Supplement). Second, making decisions with fixed critical probability thresholds leads to maximum forecast value only for a very small set of users, and using an optimisation-based approach makes better use of probabilistic forecast information (see Figure 5). Third, we showed that forecast value varies by both decision type and how expensive mitigation is for the user, highlighting the importance of calculating forecast value with the decision type which matches the real-world decision (see Figure 6). Fourth, risk aversion has a varied impact (minor to moderate) on forecast value (see Figure 7) and the degree of impact is sensitive to the decision context being evaluated. And finally, the key mechanism driving this impact is decision thresholds used relative to the damage function (see Figure 8).

Commented [RL54]: 2.29

Commented [RL55]: 2.30

Commented [RL56]: 1.35

Commented [RL57]: 2.31

6.1 Benefits of RUV over alternatives

~~1.~~ *Forecast value complements forecast verification.* Unlike forecast verification, forecast value considers the broader context within which decisions are made. This allows forecast producers, such as the Australian Bureau of Meteorology, to understand their ~~customer~~user impact by evaluating service enhancements against user decisions. ~~Determining~~Forecast verification is typically a key deciding factor when determining which method or enhancement to operationalise ~~is typically made using forecast verification as a key deciding factor~~. Quantifying the value of forecasts based on impact offers a complementary line of evidence which places the forecast user at the centre of the conversation. Because RUV encourages a dialog between the forecast producer and user to define the full decision context it may enhance communication and service adoption. For forecast users, it provides a new capability: an evidence-based approach to decide which forecast information and decision-making process will improve their outcomes. For example, the ~~Biggara~~illustrative case study in Sect. 5 indicates that subseasonal forecasts at Biggara offer better value than ~~climatology~~reference baseline in almost all cases, and that an optimisation-approach is beneficial when deciding to take early action to mitigate damages from a high flow event a few weeks ahead (see Figure 5 and Figure 6).

~~2.~~ ~~*RUV*~~*RUV is more flexible than REV.* It can model more decisions with sufficient realism than REV because it explicitly specifies decision type, risk aversion, economic model, and decision-making approach. Real-world decisions may be binary, multi-categorical, or based on continuous-flow, and using a binary model (as in REV) in all cases will provide a misleading measure of forecast value for non-binary decisions. Figure 6 shows that neglecting this would have important implications for ~~decision-makers~~users; forecasts beyond week 2 should be used for the multi-categorical and continuous-flow but not for the binary decision (when ~~$\alpha > 0.6$).~~ $\alpha > 0.6$). Similarly, neglecting the realism of other aspects of the decision may lead to other misleading conclusions. The flexibility of RUV allows the user to decide how much realism to include in the forecast value assessment depending on the information available and tailor it to the decision context.

~~3.~~ *RUV evaluates forecast value conditioned ~~upon decision-makers exposure to damages.~~on how expensive a user's mitigation is.* Unlike single-valued metrics, common in traditional forecast verification, RUV is evaluated for wide range of ~~decision-makers' exposure to damages~~users' experiences, as is shown in the value diagram (Figure 1). This offers valuable insight that would otherwise be hidden. In particular, it is useful for forecast producers who can quickly compare one forecast system to another ~~for~~over a range of ~~different~~ users with different ~~exposures to damages.~~relative expenses of mitigation (α). However, this does make it comparatively more difficult to summarise and aggregate. To assist interpretation for a single ~~forecast user~~decision-maker, it is important ~~they narrow~~the decision-maker narrows the range of ~~α which~~ α that is relevant to their decision by considering how expensive their ~~specific exposure to~~mitigation of damages is.

Formatted: Normal, No bullets or numbering

Commented [RL58]: 1.36 (and other places paragraph numbering is removed in section 6)

Field Code Changed

Field Code Changed

Commented [RL59]: 2.32

Field Code Changed

6.2 Implications of case study ~~results~~experiments

~~1.~~ *Optimisation based decision-making is better than fixed critical probability thresholds when using probabilistic forecasts.* Figure 5 demonstrates that a specific critical probability threshold will only be optimal for a specific ~~exposure to damages (α)~~value of α and suboptimal for all other values. When a ~~decision-maker~~user is choosing between using the forecast or the ~~climatological~~ reference baseline, they may choose incorrectly if their critical probability threshold is not aligned with their ~~exposure to damages.~~relative expense of mitigation. This incorrect choice will be due to a deficiency in the threshold-approach to decision-making rather than the forecast information. This RUV based finding is well supported by the REV literature ~~(Richardson, 2000; Wilks, 2001; Zhu et al., 2002; Roulin, 2007)~~(Richardson, 2000; Wilks, 2001; Zhu et al., 2002; Roulin, 2007). A perfect critical probability threshold is typically used with REV (Figure 4), unfortunately this is not possible to achieve in practice and the quantified value is unrealistically high. Matte et al. ~~(2017)~~(2017) introduced an optimisation-approach and we extended it here to further evaluate the impact on forecast value. This flexible approach makes best use of the forecast information available and for risk neutral users is equivalent to the threshold-approach when the threshold is set equal to the ~~decision-makers' exposure to damages~~user's relative expense of mitigation, ~~α~~ α (Figure 5). When forecasts are reliable this method yields value ~~which~~that is very close to the maximum possible, and forecast users may consider adopting this alternative approach for daily operational decisions. ~~A~~For this approach to be adopted for operational decision making, a Decision Support System would be required ~~so~~to calculate the optimal amount to spend on preventative mitigation ~~can be calculated~~ each time a new forecast is issued. This implies a suitable economic model is available for the decision and can be used for this calculation.

~~2.~~ *Forecast information is more valuable for risk averse users making high-stakes decisions.* Figure 7 (middle row) demonstrates that for a given forecast, a more risk averse ~~decision-maker~~user spends more to mitigate a potential damaging event than a less risk averse decision maker, all else being equal. This behaviour is consistent with their preference for risk aversion because it leads to a more certain result, with the net outcome equal to the spend amount whether the event occurs or not. There is a large difference in impact of risk aversion for the different decision types however and Figure 8 ~~summaries the~~summarises the findings of an investigation into this. Decision thresholds corresponding to very high flows lead to a larger whole impact. This finding explains why risk aversion has a large impact for the continuous-flow decision, spanning the whole regime, and a negligible impact for the binary decision with a single moderately high decision threshold. It suggests that for a risk averse user making a high stakes decision, forecasts become increasingly more valuable as the potential damages become larger. It may also explain apparently contradictory findings on the impact of risk aversion in the literature. Matte et al. ~~(2017)~~(2017) assessed the impact of risk aversion on a multi-categorical decision (using overspend and utility metrics) and found it had a moderate impact (similar to the multi-categorical decisions shown in Figure 7e and Figure 7h). Their study used 12 uniformly spaced flow divisions over a high flow range and a damage function based on empirical flood studies, whereas this study used 4 widely spaced thresholds over a similar high flow range. A recent study by Lala et al. ~~(2021)~~(2021) found minor impacts from risk aversion for binary cost-loss decisions with extreme rainfall forecasts using the

Formatted: Normal, No bullets or numbering

Field Code Changed

Field Code Changed

Commented [RL60]: 2.33

Commented [RL61]: 1.37

same expected utility maximisation framework from Matte et al. ~~(2017)~~(2017) and found a similar impact to Figure 7a. An alternative argument using reasoning from decision theory suggests that for a given risk premium the impact should be larger when decision thresholds are closer together ~~(Mas-Colell, 1995).~~(Mas-Colell, 1995). However, when investigated we found no evidence to support this~~.~~ for our case study experiments. Further research to better characterise the response for different decision contexts would be useful because the impact is modulated by both the decision thresholds and the specific damage function, consideration of the inherent sampling error introduced for extreme events would also be useful.

6.3 Limitations and future work

~~Future work on the RUV metric will focus on the following aspects:~~

~~1.~~ *Exploring the impact of alternative damage functions, economic models, ~~and~~ utility functions, and reference baselines on forecast value.* This ~~manuscript~~study focused on the impact of alternative decision types and risk aversion, and a comparative study of RUV and REV. The foundation in Expected Utility Theory allows us to model more decisions more realistically than REV, but it requires more information. When this information is unavailable or uncertain the user ~~needs~~is required to ~~apply~~make assumptions, but it is not always clear ~~what the~~how to best ~~strategy to take is~~do this. One strategy is to model all decisions as binary, cost-loss, and risk neutral and effectively convert RUV to REV. This study explores the implications of relaxing some, but not all, of those assumptions but is limited to an analysis at a single ~~case study~~forecast location. In particular, the damage function used was parameterised to simplify the introduction of RUV, facilitate comparison with REV, and highlight important implications for future studies. Further work will consider the impact of alternative damage functions and economic models tailored to other decision contexts. More descriptive economic models than cost-loss will be essential to consider decisions which involve non-economic intangible externalities like social, cultural, and ecological factors ~~(Jackson and Moggridge, 2019; Expósito et al., 2020).~~ (Jackson and Moggridge, 2019; Expósito et al., 2020). Future studies which consider these impacts may be required address unresolved findings in our study, such as the dependence of forecast value on lead-time (see Section 5.3).

~~2. *Expected Utility Theory approximates decision-making and contemporary frameworks may enhance the capability of RUV to model real-world decisions.* There is general agreement, and a substantial body of evidence, that Expected Utility Theory does not adequately describe individual choice (Kahneman and Tversky, 1979; Harless and Camerer, 1994). Many alternative models have been proposed which address these violations, such as Cumulative Prospect Theory (Tversky and Kahneman, 1992). Future work will consider whether quantifying forecast value using a foundation built on a better model of decision-making changes the conclusions reached.~~

~~3. *Exploring the relationship between forecast value and forecast skill.* Roebber and Bosart (1996) found that statistical performance metrics were poor at predicting the cost-loss value of meteorological forecasts for several real-world decisions. The relationship was impacted by the decision-maker's α value, and when in aggregate, the distribution of α over all users. Using a real-time optimisation system to manage reservoir operations Peñuela et al. (2020) quantified forecast value through improvement in pumping costs and resource availability relative to a~~

Formatted: Normal, Space After: 0 pt, No bullets or numbering

Commented [RL62]: 1.31 (noted in 2.4b)

~~baseline. They found a relationship between forecast value and CRPS skill score mediated by user priorities and hydrological conditions. Although a relationship exists it is clearly mediated by the characteristics of the decision and decision-maker and in many cases forecast skill is not a good proxy for forecast value (Murphy and Ehrendorfer, 1987; Wilks and Hamill, 1995; Roebber and Bosart, 1996; Roulin, 2007; Peñuela et al., 2020). Exploring this relationship is of interest because the decision and decision-maker characteristics are made explicit in RUV. Converting RUV to a single-value metric by placing assumptions on the distribution of α could assist and additionally allow its use as an objective function for model calibration or as a summary statistic, Wilks (2001) considers this using REV.~~

Tailoring the evaluation of forecast value to real-world decisions. For practical applications of RUV it is advisable to *calibrate* the damage function, decision thresholds, economic model, decision-making approach, and reference baseline to the real-world experience of the decision-makers. This *calibration* will ensure the resulting forecast value is tailored to the specific decision context and will likely lead to more user trust in the results, and subsequently more appropriate use of forecast information. While the reference baseline (fixed average climatology) used in this study enabled a direct comparison of RUV with REV, we would recommend comparison against more relevant baseline forecasts for practical applications (e.g., information currently used to inform the decision being assessed).

Expected Utility Theory approximates actual decision-making and contemporary frameworks may enhance the capability of RUV to model real-world decisions. There is general agreement, and a substantial body of evidence, that Expected Utility Theory does not adequately describe individual choice (Kahneman and Tversky, 1979; Harless and Camerer, 1994). Many alternative models have been proposed which address these violations, such as Cumulative Prospect Theory (Tversky and Kahneman, 1992). Future work could consider whether quantifying forecast value using a foundation built on a better model of decision-making changes the conclusions reached. Additionally, the cost-loss economic model used in this study implies that mitigation is preventative action to minimise forecast losses, with each forecast lead-time and forecast update treated independently of all others. Alternative economic models and decision-making frameworks may be required to explore more realistic forms of mitigation which consider temporal dependence (see Matte et al. (2017) for an approach).

Exploring the relationship between forecast value and forecast skill. Roebber and Bosart (1996) found that statistical performance metrics were poor at predicting the cost-loss value of meteorological forecasts for several real-world decisions. The relationship was impacted by the user's α value, and when in aggregate, the distribution of α over all users. Using a real-time optimisation system to manage reservoir operations Peñuela et al. (2020) quantified forecast value through improvement in pumping costs and resource availability relative to a baseline. They found a relationship between forecast value and CRPS skill score mediated by user priorities and hydrological conditions. Although a relationship exists it is clearly mediated by the characteristics of the decision and user and in many cases forecast skill is not a good proxy for forecast value (Murphy and Ehrendorfer, 1987; Wilks and Hamill, 1995; Roebber and Bosart, 1996; Roulin, 2007; Peñuela et al., 2020).

Commented [RL63]: 2.26

Commented [RL64]: 2.4d

Commented [RL65]: 2.34 (and other places)

Commented [RL66]: 2.33

Field Code Changed

Field Code Changed

Exploring this relationship is of interest because the decision and user characteristics are made explicit in RUV. Converting RUV to a single-value metric by placing assumptions on the distribution of α could assist and additionally allow its use as an objective function for model calibration or as a summary statistic, Wilks (2001) considers this using REV. The forecast value results of our illustrative case study are likely to be sensitive to flow characteristics and forecast uncertainty of our selected location. Future work will evaluate the value of streamflow forecast over different hydroclimatic conditions. Additionally, forecast skill (and reliability) is impacted by a forecast model's ability to reproduce seasonality and antecedent conditions. Although these are modelled well by the system used in this study (McInerney et al., 2020), their impact on forecast value was not considered in our sensitivity analysis. A future study assessing how RUV is impacted when models fail to reproduce seasonality, antecedent conditions, and other features would be a useful contribution to the field. The impact of seasonality and antecedent conditions on forecast value has not been considered in our sensitivity analysis and a future study assessing how RUV is impacted by them would be a useful contribution.

Field Code Changed

Commented [RL67]: 1.23

Commented [RL68]: 1.32 (noted in 2.4c)

~~7~~ 7 Conclusions

Formatted: Bullets and Numbering

~~Probabilistic~~Forecast value methods aim to quantify the potential benefits that probabilistic forecasts have ~~the potential to benefit~~for water-sensitive decisions, such as operational water resource management and emergency warning services~~, but to date their value for decision-making has not been established. Forecast value methods attempt to quantify this potential.~~. However, the most ~~commonly~~ used ~~existing~~ method to evaluate forecast value, Relative Economic Value (REV), is only suitable for specific decisions. REV is unsuitable for many real-world decisions and when applied may lead to misleading conclusions on when to use forecasts. This manuscript introduces the RUV metric, which has the same interpretation as the commonly used REV metric, but is more flexible and can be applied to a far wider range of ~~decisions.~~decision contexts. This is because many aspects of the decision-making process can be incorporated into RUV by the user and adjusted to match real-world decisions. These include the economic model, damage function, decision type, ~~and decision-maker characteristics and preferences, such as~~ user risk aversion and ~~exposure to damages. Importantly, we show that REV can be considered a special case~~relative expense of ~~the more general RUV, when applying specific restrictive assumptions~~mitigation.

Commented [RL69]: 2.2 & 2.4a

~~A case study demonstrates that subseasonal streamflow forecasts should be preferred over a reference climatology forecast for all lead times studied (max 30 days) and almost all decision-makers regardless of their risk aversion. This positive forecast value is robust to changes in decision-maker characteristics, decision types (binary, multi-categorical, and continuous flow), and decision-making approaches. However, beyond the second week, RUV indicates that decision-makers who are highly exposed to damages should use the reference climatology rather than the forecasts for the binary decision. This is not the case for the multi-categorical and continuous flow decision however, where forecasts should be preferred. In this study, risk aversion is found to have a larger impact for multi-categorical and continuous flow decisions than for binary decisions. However, this difference in impact is found to be a result of the specific decision thresholds used relative to the damage function rather than the decision type itself. With probabilistic forecasts, decisions are commonly made by first applying a~~

~~fixed critical probability threshold. We find that this fixed threshold approach to decision-making leads to sub-optimal use of the forecast information. Alternatively, an optimisation approach which finds the ideal amount to spend on each decision leads to the best use of the forecasts. This difference suggests the importance of modelling the real-world decision-making approach when quantifying forecast value. RUV was used to model both decision-making approaches in this study.~~ An illustrative case study using probabilistic subseasonal streamflow forecasts in a practically significant catchment in the Southern Murray-Darling Basin of Australia was used to compare the REV and RUV metrics under a range of decision contexts. The key findings from this case study were:

1. REV can be considered a special case of the more general RUV method.
2. Making decisions using an optimisation-based approach which uses the whole forecast distribution to determine the amount spent on mitigation makes better use of probabilistic forecast information than using a threshold-based approach with fixed critical probability thresholds.
3. Forecast value depends on the decision type and hence, it can be critically important to use a decision-type that matches the real-world decision
4. Risk averse users gain more value from forecasts than risk neutral users, but the impact can vary from minor to moderate depending on the decision context.
5. Impact of risk aversion on forecast value is mediated by how large the potential damages are for a given decision.

Findings 3-5 were generally sensitive to the user's relative expense of mitigation. For example, the impact of the decision-type was more pronounced for users with higher relative expenses of mitigation ($\alpha > 0.6$). In this case, for lead-times longer than 1 week, forecast value from RUV of a binary decision was significantly lower than for multi-categorical or continuous-flow decisions. As REV is limited to binary decisions, a user making a multi-categorical or continuous-flow decision, could be misled by the REV outcomes and consider not using the forecasts when they actually have significant value as demonstrated by RUV.

This manuscript focuses on the introduction of RUV and an exploration of its sensitivity to some aspects of decision context. Therefore, several future research directions for RUV are discussed including (i) exploring sensitivity of forecast value to more aspects of decision context, (ii) tailoring forecast value to real-world decisions, (iii) assessing alternative frameworks for modelling decision-making, and (iv) exploring the relationship between forecast value and forecast skill.

RUV presents an opportunity to tailor forecasts and their assessment to the specific decisions, decision-making approach, characteristics, preferences, and economics of the ~~decision-maker.~~user. It is hoped that this capability ~~may~~will encourage the assessment of forecast systems through the lens of ~~customer~~user benefit and be seen as a complement to forecast verification. This may lead to increased adoption of forecasts through deeper dialog and understanding, and ultimately to improved water resource management decisions.

Field Code Changed

Commented [RL70]: 2.3 & 2.4a & 2.25

Formatted: Font: +Headings (Times New Roman)

~~Appendix A~~ Appendix A Proof of equivalence of REV and RUV under specific assumptions

This section demonstrates the equivalence of the REV metric as detailed in Eq. (2) and the RUV metric introduced in Sect. 3 when 5 assumptions are applied to the decision context. A complete derivation is included in the Supplement.

In a cost-loss decision problem the two relevant states are "flow above" and "flow below" a decision threshold ~~Q_d~~ Q_d:

$$\begin{array}{ll} m = \text{above} & \text{if } Q_t \geq Q_d \\ m = \text{below} & \text{if } Q_t < Q_d \end{array} \qquad \begin{array}{ll} m = \text{above} & \text{if } Q_t \geq Q_d \\ m = \text{below} & \text{if } Q_t < Q_d \end{array} \tag{12}$$

Assumption 1: A step damage function with binary values of 0 and L is used to specify the losses above and below the decision threshold for all timesteps,

$$d(m) = \begin{cases} L & \text{when } m = \text{above} \\ 0 & \text{when } m = \text{below} \end{cases} \qquad d(m;L) = \begin{cases} L & \text{when } m = \text{above} \\ 0 & \text{when } m = \text{below} \end{cases} \tag{13}$$

To calculate the net outcome when action is taken to mitigate the loss, we substitute Eq. (7) and (13) into Eq. (6), which leads to the following net outcomes for the two states.

$$\begin{array}{ll} E_{t,above} = \min\left(\beta \cdot C_t, L_t\right) - L_t - C_t & \\ E_{t,below} = -C_t & \text{since } \beta \cdot C_t > 0 \end{array} \qquad \begin{array}{ll} E_{t,above} = \min\left(\beta \cdot C_t, L\right) - L - C_t & \\ E_{t,below} = -C_t & \text{since } \beta \cdot C_t > 0 \end{array} \tag{14}$$

Assumption 2: Linear utility function is assumed which implies no aversion to risk,

$$\mu(E) = E \qquad \mu(E) = E \tag{15}$$

Substituting Eq. (14) into Eq. (4), applying the linear utility function assumption, and simplifying for only two possible states using ~~p~~ p_t, the forecast probability of flow above the flow threshold, at time t, leads to.

$$\begin{aligned} U(\tilde{E}_t) &= p_t \cdot E_t^{above} + (1 - p_t) \cdot E_t^{below} \\ &= p_t \cdot \left[\min\left(\beta \cdot C_t, L_t\right) - L_t - C_t\right] + (1 - p_t) \cdot \left[-C_t\right] \\ U(E_t) &= p_t \cdot E_{t,above} + (1 - p_t) \cdot E_{t,below} \\ &= p_t \cdot \left[\min\left(\beta \cdot C_t, L\right) - L - C_t\right] + (1 - p_t) \cdot \left[-C_t\right] \end{aligned} \tag{16}$$

Assumption 3: Probability of flow above the threshold will always be either 1 or 0,

$$p_t \in \{0,1\} \qquad p_t \in \{0,1\} \tag{17}$$

~~We~~Using these assumptions and noting that the total loses at each timestep are fixed and consist of avoided and un-avoided components $L = L_t = L_t^a + L_t^u$ we can now determine the single timestep ex ante utility for the four possible outcomes~~:~~. The

Formatted: Bullets and Numbering

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

four possible outcomes are composed of; event is forecast ~~probability is 1 or 0,~~ to occur ($p_t = 1$) or not occur ($p_t = 0$), and an action has therefore been taken or not, leading to Table 4.

Table 4: Ex ante utility values for a time-step of Expected Utility Theory with REV assumptions

	~~$p = 1$~~ Event forecast to occur	~~$p = 0$~~ Event not forecast to occur
Action taken ~~$C_t \neq 0$~~ $C_t \neq 0$	~~$-(C_t + L_t^u)$~~ $-(C_t + L_t^u)$	~~$-C_t$~~ $-C_t$
Action not taken ~~$C_t = 0$~~ $C_t = 0$	~~$-L_t$~~ $-L$	~~0~~ 0

Applying Eq. (8) to Eq. (16) will lead to an optimal amount ~~C_t^π~~ $\bar{C}_t$ to spend on the mitigating action for each timestep. By considering that the forecast probability is always either 1 or 0 due to assumption 3 and that all costs and losses are positive values we can derive that for any timestep the cost will be either ~~$C_t^\pi = 0$ when $p = 0$ or $C_t^\pi = \frac{L_t}{\beta}$ when $p = 1$.~~ $\bar{C}_t = 0$ when $p_t = 0$ or $\bar{C}_t = \frac{L}{\beta}$ when $p_t = 1$, see the supplement for a full derivation.

The ex post utility for each timestep, shown in Table 5, can be found by substituting these optimal costs back into the elements of Table 4, and letting the probability be conditioned on the state of observed flow above the threshold.

Table 5: Ex post utility values for a time-step of Expected Utility Theory with REV assumptions

	Event occurred ~~$p_t = 1$~~	Event did not occur ~~$p_t = 0$~~
Action taken ~~$C_t = \frac{L_t}{\beta}$~~ $C_t = \frac{L}{\beta}$	~~$-\left(\frac{L_t}{\beta} + L_t^u\right)$~~ $-\left(\frac{L}{\beta} + L_t^u\right)$	~~$\frac{L_t}{\beta}$~~ $-\frac{L}{\beta}$
Action not taken ~~$C_t = 0$~~ $C_t = 0$	~~$-L_t$~~ $-L$	~~0~~ 0

A contingency table is now used with Table 5 to determine each term of the RUV metric.

Assumption 4: The frequency of the binary decision event ~~$\overline{o}$~~ $\bar{o}$ is used for the reference baseline.

Field Code Changed
Field Code Changed
Commented [RL71]: 1.38
Commented [RL72]: 1.38 (and table S2)
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Commented [RL73]: 1.38
Field Code Changed
Commented [RL74]: 1.38
Field Code Changed
Field Code Changed
Field Code Changed
Commented [RL75]: 1.38 (and Table S3)
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed

This leads to the following expected ex post utility for reference baseline information

~~$\mathop{\mathbb{E}}_{t\in T}\left[\Upsilon\left(E_t^r\right)\right] = \min\left\{\frac{L_t}{\beta}, \bar{o}L_t^a\right\} - \bar{o}L_t^u$~~ $\mathop{\mathbb{E}}_{t\in T}\left[\Upsilon\left(E_t^r\right)\right] = -\min\left\{\frac{L}{\beta}, \bar{o}L_t^a\right\} - \bar{o}L_t^u$ (18)

Field Code Changed

Expected ex post utility for perfect information is

~~$\mathop{\mathbb{E}}_{t\in T}\left[\Upsilon\left(E_t^p\right)\right] = \bar{o}\left(\frac{L_t}{\beta} + L_t^u\right)$~~ $\mathop{\mathbb{E}}_{t\in T}\left[\Upsilon\left(E_t^p\right)\right] = -\bar{o}\left(\frac{L}{\beta} + L_t^u\right)$ (19)

Field Code Changed

Expected ex post utility for forecast information is

~~$\mathop{\mathbb{E}}_{t\in T}\left[\Upsilon\left(E_t^f\right)\right] = (h+f)\frac{L_t}{\beta} - \bar{o}L_t^u - mL_t^a$~~ $\mathop{\mathbb{E}}_{t\in T}\left[\Upsilon\left(E_t^f\right)\right] = -(h+f)\frac{L}{\beta} - \bar{o}L_t^u - mL_t^a$ (20)

Field Code Changed

where ~~h~~ h is the hit rate, ~~m~~ m is the miss rate, and ~~f~~ f is the false alarm rate from the contingency table.

Field Code Changed

Field Code Changed

Field Code Changed

Assumption 5: At each timestep the avoided losses are equal to the total possible losses.

~~$L_t = L_t^a \quad \text{for } t \in T$~~ $L_t^a = L \quad \text{for } t \in T$ (21)

Field Code Changed

Substituting Eq. (18), (19), and (20) into Eq. (10), applying assumption 5, and noting the relationship ~~$\beta = \frac{1}{\alpha}$~~ $\beta = \frac{1}{\alpha}$ leads to

~~$RUV = \frac{\min(\alpha,\bar{o}) - (h+f)\alpha - m}{\min(\alpha,\bar{o}) - \bar{o}\alpha}$~~ $\text{RUV} = \frac{\min(\alpha,\bar{o}) - (h+f)\alpha - m}{\min(\alpha,\bar{o}) - \bar{o}\alpha}$ (22)

Field Code Changed

which is identical to the definition of the REV metric in Eq. (2).

Code availability

The code used for this work will be released, along with a follow up publication, as a software library which can be used by researchers and industry to quantify forecast value using RUV. Please contact the corresponding author to register interest in beta testing access.

Data availability

~~A companion dataset for this work is available at: https://doi.org/XXX. This contains the input streamflow forecasts, output forecast value results, and generated figures. The software library used to generate the forecast value results is not included in this dataset because it will be released with a follow up publication.~~A companion dataset for this work is available at: https://doi.org/10.25909/19153055 (Laugesen et al., 2022). This contains the input streamflow forecasts, output forecast value results, and high resolution figures.

Commented [RL76]: 1.39

Author contributions

Richard Laugesen led the conceptualisation, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualisation, writing – original draft preparation, and writing – review & editing. Mark Thyer supported funding acquisition, investigation, methodology, project administration, supervision, visualisation, and writing – review & editing. David McInerney supported formal analysis, methodology, resources, visualisation, and writing – review & editing. Dmitri Kavetski supported methodology, visualisation, and writing – review & editing.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements.

This work was conducted on the traditional lands of the Ngunnawal people and Kaurna people. We acknowledge their continuing custodianship of these lands and the rivers that flow through them, and pay our respects to their elders, past ~~and~~, present~~.~~, and emerging. We also acknowledge the Jaitmatang and Ngarigo people, traditional custodians of the ~~catchments and rivers~~upper Murray River catchment used in this study. The authors thank two anonymous journal reviewers, Beth Ebert, Michael Foley, and Prasantha Hapuarachchi for their review of this paper and thoughtful discussions on the method, and Jacqui Hickey for her formative discussions on the use of forecasts for operational decision making at the MDBA and encouragement to pursue this topic. Richard Laugesen is grateful to the Bureau of Meteorology for their generous support of his research, particularly Narendra Tuteja, Alex Cornish, and Adam Smith for seeing the value of this ~~approach~~innovation. This work was supported through an Australian Government Research Training Program Scholarship and with supercomputing resources provided by the Phoenix HPC service at the University of Adelaide.

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, Mon. Wea. Rev., 141, 3462–3476, https://doi.org/10.1175/MWR-D-12-00206.1, 2013.

~~Abaza, M., Anctil, F., Fortin, V.,~~Anghileri, D., Monhart, S., Zhou, C., Bogner, K., Castelletti, A., Burlando, P., and ~~Turcotte~~Zappa, M.: The Value of Subseasonal Hydrometeorological Forecasts to Hydropower Operations: How Much Does Preprocessing Matter?, Water Resources Research, 55, 10159–10178, https://doi.org/10.1029/2019WR025280, 2019.

An-Vo, D.-A., Mushtaq, S., Reardon-Smith, K., Kouadio, L., Attard, S., Cobon, D., and Stone, R.: ~~Sequential streamflow assimilation for short-term hydrological ensemble~~Value of seasonal forecasting~~,~~ for sugarcane farm irrigation planning, European Journal of ~~Hydrology, 519, 2692–2706~~Agronomy, 104, 37–48, https://doi.org/10.1016/j.~~jhydrol.2014.08.038, 2014~~eja.2019.01.005, 2019.

Babcock, B. A., Choi, E. K., and Feinerman, E.: Risk and probability premiums for CARA utility functions, Journal of Agricultural and Resource Economics, 18, 17–24, https://doi.org/10.22004/ag.econ.30810, 1993.

Bennett, J. C., Robertson, D. E., Wang, Q. J., Li, M., and Perraud, J.-M.: Propagating reliable estimates of hydrological forecast uncertainty to many lead times, Journal of Hydrology, 603, 126798, https://doi.org/10.1016/j.jhydrol.2021.126798, 2021.

~~Bergh, J. V. den and Roulin, E.: Hydrological ensemble prediction and verification for the Meuse and Scheldt basins, 11, 64–71, https://doi.org/10.1002/asl.250, 2010.~~

Bischiniotis, K., van den Hurk, B., Coughlan de Perez, E., Veldkamp, T., Nobre, G. G., and Aerts, J.: Assessing time, cost and quality trade-offs in forecast-based action for floods, International Journal of Disaster Risk Reduction, 40, 101252, https://doi.org/10.1016/j.ijdrr.2019.101252, 2019.

Bogner, K., ~~Cloke, H. L., Pappenberger, F., de Roo, A.,~~Liechti, K., and Zappa, M.: Post-Processing of Stream Flows in Switzerland with an Emphasis on Low Flows and ~~Thielen, J.: Improving the evaluation of hydrological multi-model forecast performance in the Upper Danube Catchment, 10, 1–12~~Floods, Water, 8, 115, https://doi.org/10.~~1080/15715124.2011.625359, 2012~~3390/w8040115, 2016.

Cantonati, M., Poikane, S., Pringle, C. M., Stevens, L. E., Turak, E., Heino, J., Richardson, J. S., Bolpagni, R., Borrini, A., Cid, N., Čtvrtlíková, M., Galassi, D. M. P., Hájek, M., Hawes, I., Levkov, Z., Naselli-Flores, L., Saber, A. A., Cicco, M. D., Fiasca, B., Hamilton, P. B., Kubečka, J., Segadelli, S., and Znachor, P.: Characteristics, Main Impacts, and Stewardship of Natural and Artificial Freshwater Environments: Consequences for Biodiversity Conservation, Water, 12, 260, https://doi.org/10.3390/w12010260, 2020.

Carr, R. H., Semmens, K., Montz, B., and Maxfield, K.: Improving the Use of Hydrologic Probabilistic and Deterministic Information in Decision-Making, Bulletin of the American Meteorological Society, 102, E1878–E1896, https://doi.org/10.1175/BAMS-D-21-0019.1, 2021.

Cassagnole, M., Ramos, M.-H., Zalachori, I., Thirel, G., Garçon, R., Gailhard, J., and Ouillon, T.: Impact of the quality of hydrological forecasts on the management and revenue of hydroelectric reservoirs – a conceptual approach, Hydrology and Earth System Sciences, 25, 1033–1052, https://doi.org/10.5194/hess-25-1033-2021, 2021.

Cloke, H. L. and Pappenberger, F.: Ensemble flood forecasting: A review, Journal of Hydrology, 375, 613–626, https://doi.org/10.1016/j.jhydrol.2009.06.005, 2009.

Dorrington, J., Finney, I., Palmer, T., and Weisheimer, A.: Beyond skill scores: exploring sub-seasonal forecast value through a case-study of French month-ahead energy prediction, Quarterly Journal of the Royal Meteorological Society, 146, 3623–3637, https://doi.org/10.1002/qj.3863, 2020.

Duan, Q. editor, Pappenberger, F. editor, Wood, A. editor, Cloke, H. L. editor, and Schaake, J. C. editor: Handbook of Hydrometeorological Ensemble Forecasting edited by Qingyun Duan, Florian Pappenberger, Andy Wood, Hannah L. Cloke, John C. Schaake, Berlin, Heidelberg : Springer Berlin Heidelberg : Imprint: Springer, 2019.

Expósito, A., Beier, F., and Berbel, J.: Hydro-Economic Modelling for Water-Policy Assessment Under Climate Change at a River Basin Scale: A Review, Water, 12, 1559, https://doi.org/10.3390/w12061559, 2020.

Foley, M. and Loveday, N.: Comparison of Single-Valued Forecasts in a User-Oriented Framework, Weather and Forecasting, 35, 1067–1080, https://doi.org/10.1175/WAF-D-19-0248.1, 2020.

~~Fundel, F., Jörg-Hess, S., and Zappa, M.: Monthly hydrometeorological ensemble prediction of streamflow droughts and corresponding drought indices, Hydrol. Earth Syst. Sci., 17, 395–407, https://doi.org/10.5194/hess-17-395-2013, 2013.~~

Fundel, V. J., Fleischhut, N., Herzog, S. M., Göber, M., and Hagedorn, R.: Promoting the use of probabilistic weather forecasts through a dialogue between scientists, developers and end-users, Quarterly Journal of the Royal Meteorological Society, 145, 210–231, https://doi.org/10.1002/qj.3482, 2019.

Grafton, R. Q. and Wheeler, S. A.: Economics of Water Recovery in the Murray-Darling Basin, Australia, Annu. Rev. Resour. Econ., 10, 487–510, https://doi.org/10.1146/annurev-resource-100517-023039, 2018.

Harless, D. W. and Camerer, C. F.: The Predictive Utility of Generalized Expected Utility Theories, Econometrica, 62, 1251–1289, https://doi.org/10.2307/2951749, 1994.

Hudson et. al., D.: ACCESS-S1 The new Bureau of Meteorology multi-week to seasonal prediction system, JSHESS, 67, 132–159, https://doi.org/10.22499/3.6703.001, 2017.

Jackson, S. and Moggridge, B.: Indigenous water management, Australasian Journal of Environmental Management, 26, 193–196, https://doi.org/10.1080/14486563.2019.1661645, 2019.

Jones, D., Wang, W., and Fawcett, R. J. B.: High-quality spatial climate data-sets for Australia, Australian Meteorological and Oceanographic Journal, 58, 233–248, 2009.

Kahneman, D. and Tversky, A.: Prospect Theory: An Analysis of Decision under Risk, Econometrica, 47, 263–291, https://doi.org/10.2307/1914185, 1979.

Katz, R. W. and Lazo, J. K.: Economic Value of Weather and Climate Forecasts, Oxford University Press, https://doi.org/10.1093/oxfordhb/9780195398649.013.0021, 2011.

Katz, R. W. and Murphy, A. H. (Eds.): Economic Value of Weather and Climate Forecasts, Cambridge University Press, Cambridge, https://doi.org/10.1017/CBO9780511608278, 1997.

Laio, F. and Tamea, S.: Verification tools for probabilistic forecasts of continuous hydrological variables, Hydrol. Earth Syst. Sci., 11, 1267–1277, https://doi.org/10.5194/hess-11-1267-2007, 2007.

Lala, J., Bazo, J., Anand, V., and Block, P.: Optimizing forecast-based actions for extreme rainfall events, Climate Risk Management, 34, 100374, https://doi.org/10.1016/j.crm.2021.100374, 2021.

Laugesen, R., Thyer, M., McInerney, D., and Kavetski, D.: Supporting data for "Flexible forecast value metric suitable for a wide range of decisions: application using probabilistic subseasonal streamflow forecasts" by Laugesen et.al. (2022), https://doi.org/10.25909/19153055.v1, 2022.

Li, C., Cheng, X., Li, N., Liang, Z., Wang, Y., and Han, S.: A Three-Parameter S-Shaped Function of Flood Return Period and Damage, Advances in Meteorology, 2016, e6583906, https://doi.org/10.1155/2016/6583906, 2016a.

Li, M., Wang, Q. J., Bennett, J. C., and Robertson, D. E.: Error reduction and representation in stages (ERRIS) in hydrological modelling for ensemble streamflow forecasting, Hydrology and Earth System Sciences, 20, 3561–3579, https://doi.org/10.5194/hess-20-3561-2016, 2016b.

Lopez, A., Coughlan de Perez, E., Bazo, J., Suarez, P., van den Hurk, B., and van Aalst, M.: Bridging forecast verification and humanitarian decisions: A valuation approach for setting up action-oriented early warnings, Weather and Climate Extremes, 27, 100167, https://doi.org/10.1016/j.wace.2018.03.006, 2020.

Lucatero, D., Madsen, H., Refsgaard, J. C., Kidmose, J., and Jensen, K. H.: Seasonal streamflow forecasts in the Ahlergaarde catchment, Denmark: the effect of preprocessing and post-processing on skill and statistical consistency, Hydrology and Earth System Sciences, 22, 3601–3617, https://doi.org/10.5194/hess-22-3601-2018, 2018.

Marzban, C.: Displaying Economic Value, Weather and Forecasting, 27, 1604–1612, https://doi.org/10.1175/WAF-D-11-00138.1, 2012.

Mas-Colell, A.: Microeconomic theory, Oxford University Press, New York, xvii+981 pp., 1995.

Matte, S., Boucher, M.-A., Boucher, V., and Fortier Filion, T.-C.: Moving beyond the cost–loss ratio: economic assessment of streamflow forecasts for a risk-averse decision maker, Hydrology and Earth System Sciences, 21, 2967–2986, https://doi.org/10.5194/hess-21-2967-2017, 2017.

McInerney, D., Thyer, M., Kavetski, D., Lerat, J., and Kuczera, G.: Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors, Water Resources Research, 53, 2199–2239, https://doi.org/10.1002/2016WR019168, 2017.

McInerney, D., Thyer, M., Kavetski, D., Laugesen, R., Tuteja, N., and Kuczera, G.: Multi-temporal Hydrological Residual Error Modeling for Seamless Subseasonal Streamflow Forecasting, Water Resources Research, 56, e2019WR026979, https://doi.org/10.1029/2019WR026979, 2020.

McInerney, D., Thyer, M., Kavetski, D., Laugesen, R., Woldemeskel, F., Tuteja, N., and Kuczera, G.: Improving the Reliability of Sub-Seasonal Forecasts of High and Low Flows by Using a Flow-Dependent Nonparametric Model, Water Resources Research, 57, e2020WR029317, https://doi.org/10.1029/2020WR029317, 2021.

McInerney, D., Thyer, M., Kavetski, D., Laugesen, R., Woldemeskel, F., Tuteja, N., and Kuczera, G.: Seamless streamflow model provides forecasts at all scales from daily to monthly and matches the performance of non-seamless monthly model, Hydrology and Earth System Sciences Discussions, 1–22, https://doi.org/10.5194/hess-2021-589, 2022.

Monhart, S., Zappa, M., Spirig, C., Schär, C., and Bogner, K.: Subseasonal hydrometeorological ensemble predictions in small- and medium-sized mountainous catchments: benefits of the NWP approach, Hydrology and Earth System Sciences, 23, 493–513, https://doi.org/10.5194/hess-23-493-2019, 2019.

Murphy, A. H.: Value of climatological, categorical and probabilistic forecasts in cost-loss ratio situation, Monthly Weather Review, 105, 803–816, https://doi.org/10.1175/1520-0493(1977)105<0803:tvocca>2.0.co;2, 1977.

Murphy, A. H.: What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting, Weather and Forecasting, 8, 281–293, https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2, 1993.

Murphy, A. H. and Ehrendorfer, M.: On the Relationship between the Accuracy and Value of Forecasts in the Cost–Loss Ratio Situation, Weather and Forecasting, 2, 243–251, https://doi.org/10.1175/1520-0434(1987)002<0243:OTRBTA>2.0.CO;2, 1987.

Murray–Darling Basin Authority: Modelling assessment to determine SDL Adjustment Volume, Murray–Darling Basin Authority, Canberra, Australia, 2017.

Mylne, K. R.: Decision-making from probability forecasts based on forecast value, Meteorological Applications, 9, 307–315, https://doi.org/10.1017/s1350482702003043, 2002.

Neumann, J. V.: Theory Of Games And Economic Behavior, 1944.

Palmer, T. N.: The economic value of ensemble forecasts as a tool for risk assessment: From days to decades, Quarterly Journal of the Royal Meteorological Society, 128, 747–774, https://doi.org/10.1256/0035900021643593, 2002.

Peñuela, A., Hutton, C., and Pianosi, F.: Assessing the value of seasonal hydrological forecasts for improving water resource management: insights from a pilot application in the UK, Hydrol. Earth Syst. Sci., 24, 6059–6073, https://doi.org/10.5194/hess-24-6059-2020, 2020.

Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, Journal of Hydrology, 279, 275–289, https://doi.org/10.1016/S0022-1694(03)00225-7, 2003.

Portele, T. C., Lorenz, C., Dibrani, B., Laux, P., Bliefernicht, J., and Kunstmann, H.: Seasonal forecasts offer economic benefit for hydrological decision making in semi-arid regions, Sci Rep, 11, 10581, https://doi.org/10.1038/s41598-021-89564-y, 2021.

Pratt, J. W.: Risk Aversion in the Small and in the Large, Econometrica, 32, 122–136, https://doi.org/10.2307/1913738, 1964.

Richardson, D. S.: Skill and relative economic value of the ECMWF ensemble prediction system, Quarterly Journal of the Royal Meteorological Society, 126, 649–667, https://doi.org/10.1256/smsqj.56312, 2000.

Roebber, P. J. and Bosart, L. F.: The Complex Relationship between Forecast Skill and Forecast Value: A Real-World Analysis, Wea. Forecasting, 11, 544–559, https://doi.org/10.1175/1520-0434(1996)011<0544:TCRBFS>2.0.CO;2, 1996.

Roulin, E.: Skill and relative economic value of medium-range hydrological ensemble predictions, Hydrology and Earth System Sciences, 11, 725–737, https://doi.org/10.5194/hess-11-725-2007, 2007.

Schepen, A., Zhao, T., Wang, Q. J., and Robertson, D. E.: A Bayesian modelling method for post-processing daily sub-seasonal to seasonal rainfall forecasts from global climate models and evaluation for 12 Australian catchments, Hydrology and Earth System Sciences, 22, 1615–1628, https://doi.org/10.5194/hess-22-1615-2018, 2018.

Schmitt Quedi, E. and Mainardi Fan, F.: Sub seasonal streamflow forecast assessment at large-scale basins, Journal of Hydrology, 584, 124635, https://doi.org/10.1016/j.jhydrol.2020.124635, 2020.

Soares, M. B., Daly, M., and Dessai, S.: Assessing the value of seasonal climate forecasts for decision-making, Wiley Interdisciplinary Reviews-Climate Change, 9, https://doi.org/10.1002/wcc.523, 2018.

Tabari, H.: Climate change impact on flood and extreme precipitation increases with water availability, Sci Rep, 10, 13768, https://doi.org/10.1038/s41598-020-70816-2, 2020.

Tena, E. C. and Gómez, S. Q.: Cost-Loss Decision Models with Risk Aversion, 28, 2008.

Thiboult, A., Anctil, F., and Ramos, M. H.: How does the quantification of uncertainties affect the quality and value of flood early warning systems?, Journal of Hydrology, 551, 365–373, https://doi.org/10.1016/j.jhydrol.2017.05.014, 2017.

Thompson, J. C.: On the Operational Deficiences in Categorical Weather Forecasts, Bulletin of the American Meteorological Society, 33, 223–226, 1952.

Turner, S. W. D., Bennett, J. C., Robertson, D. E., and Galelli, S.: Complex relationship between seasonal streamflow forecast skill and value in reservoir operations, Hydrol. Earth Syst. Sci., 21, 4841–4859, https://doi.org/10.5194/hess-21-4841-2017, 2017.

Tversky, A. and Kahneman, D.: Advances in prospect theory: Cumulative representation of uncertainty, J Risk Uncertainty, 5, 297–323, https://doi.org/10.1007/BF00122574, 1992.

UNESCO (Ed.): Managing water under uncertainty and risk, UNESCO [u.a.], Paris, 780 pp., 2012.

United Nations: International UN-Water Conference. Water in the Green Economy in Practice: Towards Rio+20, 2011.

Verkade, J. S~~. and Werner, M. G. F.: Estimating the benefits of single value and probability forecasting for flood warning, 15, 3751–3765, https://doi.org/10.5194/hess-15-3751-2011, 2011.~~

~~Verkade, J. S.~~, Brown, J. D., Davids, F., Reggiani, P., and Weerts, A. H.: Estimating predictive hydrological uncertainty by dressing deterministic and ensemble forecasts; a comparison, with application to Meuse and Rhine, Journal of Hydrology, 555, 257–277, https://doi.org/10.1016/j.jhydrol.2017.10.024, 2017.

~~Weijs, S. V., Schoups, G., and Giesen, N. van de: Why hydrological predictions should be evaluated using information theory, 14, 2545–2558, https://doi.org/10.5194/hess-14-2545-2010, 2010.~~

Werner, J.: Risk Aversion, in: The New Palgrave Dictionary of Economics, Palgrave Macmillan UK, 1–6, https://doi.org/10.1057/978-1-349-95121-5_2741-1, 2008.

White, C. J., Franks, S. W., and McEvoy, D.: Using subseasonal-to-seasonal (S2S) extreme rainfall forecasts for extended-range flood prediction in Australia, in: Proceedings of the International Association of Hydrological Sciences, Changes in Flood Risk and Perception in Catchments and Cities - IAHS Symposium HS01, 26th General Assembly of the International Union of Geodesy and Geophysics, Prague, Czech Republic, 22 June–2 July 2015, 229–234, https://doi.org/10.5194/piahs-370-229-2015, 2015.

Wilks, D. S.: A skill score based on economic value for probability forecasts, Meteorological Applications, 8, 209–219, https://doi.org/10.1017/S1350482701002092, 2001.

Wilks, D. S. and Hamill, T. M.: Potential Economic Value of Ensemble-Based Surface Weather Forecasts, Monthly Weather Review, 123, 3565–3575, https://doi.org/10.1175/1520-0493(1995)123<3565:PEVOEB>2.0.CO;2, 1995.

Woldemeskel, F., McInerney, D., Lerat, J., Thyer, M., Kavetski, D., Shin, D., Tuteja, N., and Kuczera, G.: Evaluating post-processing approaches for monthly and seasonal streamflow forecasts, Hydrology and Earth System Sciences, 22, 6257–6278, https://doi.org/10.5194/hess-22-6257-2018, 2018.

Wu, W., Emerton, R., Duan, Q., Wood, A. W., Wetterhall, F., and Robertson, D. E.: Ensemble flood forecasting: Current status and future opportunities, WIREs Water, 7, e1432, https://doi.org/10.1002/wat2.1432, 2020.

Zhang, X. S., Amirthanathan, G. E., Bari, M. A., Laugesen, R. M., Shin, D., Kent, D. M., MacDonald, A. M., Turner, M. E., and Tuteja, N. K.: How streamflow has changed across Australia since the 1950s: Evidence from the network of hydrologic reference stations, Hydrology and Earth System Sciences, 20, 3947–3965, https://doi.org/10.5194/hess-20-3947-2016, 2016.

Zhu, Y. J., Toth, Z., Wobus, R., Richardson, D., and Mylne, K.: The economic value of ensemble-based weather forecasts, Bulletin of the American Meteorological Society, 83, 73-+, https://doi.org/10.1175/1520-0477(2002)083<0073:tevoeb>2.3.co;2, 2002.

Formatted: Font: +Headings (Times New Roman), 14 pt, Font color: Text 1