Decide Now or Wait for the Next Forecast? Testing a Decision Framework Using Real Forecasts and Observations

STEPHEN JEWSON, a SEBASTIAN SCHE, b,c,d AND GABRIELE MESSORI, c,d,e,f

a London, United Kingdom
b Know-Center GmbH, Graz, Austria
c Department of Meteorology, Stockholm University, Stockholm, Sweden
d Bolin Centre for Climate Research, Stockholm University, Stockholm, Sweden
e Department of Earth Sciences, Uppsala University, Uppsala, Sweden
f Centre of Natural Hazards and Disaster Science, Uppsala University, Uppsala, Sweden

(Manuscript received 27 November 2020, in final form 3 March 2021)

ABSTRACT: Users of meteorological forecasts are often faced with the question of whether to make a decision now, on the basis of the current forecast, or to wait for the next and, it is hoped, more accurate forecast before making the decision. Following previous authors, we analyze this question as an extension of the well-known cost–loss model. Within this extended cost–loss model, the question of whether to decide now or to wait depends on two specific aspects of the forecast, both of which involve probabilities of probabilities. For the special case of weather and climate forecasts in the form of normal distributions, we derive a simple simulation algorithm, and equivalent analytical expressions, for calculating these two probabilities. We apply the algorithm to forecasts of temperature and find that the algorithm leads to better decisions in most cases relative to three simpler alternative decision-making schemes, in both a simulated context and when we use reforecasts, surface observations, and rigorous out-of-sample validation of the decisions. To the best of our knowledge, this is the first time that a dynamic multistage decision algorithm has been demonstrated to work using real weather observations. Our results have implications for the additional kinds of information that forecasters of weather and climate could produce to facilitate good decision-making on the basis of their forecasts.

SIGNIFICANCE STATEMENT: Users of weather and climate forecasts often find themselves in a situation in which they could make a decision now or they could wait for the next forecast before they make the decision. Building on previous research, we derive a simple algorithm for approaching this situation in a logical way. Our algorithm can be used in many common forecasting situations, and we demonstrate that the algorithm works by testing it using forecasts and real temperature measurements.

KEYWORDS: Forecast verification/skill; Probability forecasts/models/distribution; Decision support

1. Introduction

The use of forecast probabilities to make decisions has been studied extensively in the atmospheric sciences. Many studies have investigated decision-making using the single-stage cost–loss model (e.g., Murphy 1969; Kernan 1975; Katz and Murphy 1997; Buizza 2001; Richardson 2001; Roulin 2007). There have also been a number of generalizations of the single-stage cost–loss model (Murphy 1985; Murphy et al. 1985; Epstein and Murphy 1987; Murphy and Ye 1990; Wilks 1991; Wilks et al. 1993; Katz 1993; Wilks and Wolfe 1998; Regnier and Harr 2006; Roulin 2007; McLay 2011; Tenan and Gomez 2011; Matte et al. 2017). A review is given in Wilks (2014). The single-stage cost–loss model is used as an idealized model for making types of decisions that are analogous in terms of logical structure to the binary decision of whether to cancel an event based on a forecast. In this simple model there are just two possible weather outcomes (which we will refer to as good or bad weather), with predicted probabilities, and a single binary decision that needs to be taken based on those probabilities
introduced by Richard Bellman in the 1950s (Bellman 1957). Dynamic programming is used in many fields other than weather and climate; for example, economists have used it to study whether investors should buy a stock today or wait to see how the stock price evolves as new information becomes available. Dynamic programming is often considered as a part of the larger field of operations research (Ventura 2019).

These studies lay out a body of theory for multistage dynamic decision-making, for various different decision situations, and test the theory in various ways. However, perhaps with the exception of the weather roulette game described in Hagedorn and Smith (2009) and Terrado et al. (2019), there have apparently been no studies on decision-making that have attempted to validate decisions with real weather observations. As a result, it is not possible to assess the extent to which the decision-making algorithms that these studies describe would work better than alternative simpler decision algorithms in a real situation, given the extra uncertainties that real forecasts and real observations introduce.

In this article, we reconsider the two-stage dynamic decision-making problem discussed by Murphy and Ye (1990) and Regnier and Harr (2006): should we decide now, or should we wait for the next forecast? The technical approach we take to answering this question is closely related to the methods described in the studies cited above, with some small differences. For instance, our model follows Murphy and Ye (1990), Regnier and Harr (2006), and McLay (2011) and differs from those of Murphy et al. (1985) and Epstein and Murphy (1987) and the example in Wilks (1991) in that it allows the costs in the cost–loss model to vary in time. Variations in costs are often an essential factor in deciding whether to wait for the next forecast. Our model differs from Regnier and Harr (2006) and McLay (2011) in that it involves temperature. The most important novel aspects of our analysis are that (i) we consider normally distributed forecasts, and for this special case we are able to derive an algorithm for making decisions that is much simpler than the Markov chain modeling of transition probabilities described in Wilks (1991), Regnier and Harr (2006), and McLay (2011). The algorithm requires nothing more than a knowledge of the root mean squared error performance of the forecast system, and (ii) we apply our decision algorithm to real forecast data and validate the decisions it generates against real weather observations, in a rigorous out-of-sample way. By doing so, we address the crucial question of whether the algorithm for decision-making that we describe could really be beneficial if applied in practice.

We now illustrate some of the aspects of the decide now or wait for the next forecast decision, using a simple example. An event is planned for Saturday. If the weather conditions at the start of the event are unsuitable then the event will have to be canceled, leading to various expenses, known as the “loss” in the cost–loss framework. Daily weather forecasts are available in the run-up to the event and are used by the event organizer to decide whether to cancel in advance. Canceling on Thursday leads to only small cancellation charges, while canceling on Friday leads to larger charges. Both sets of cancellation charges are lower than the potential loss due to last-minute cancellation on Saturday, and this leads to a nuanced set of decisions around whether to cancel on Thursday, Friday, or not at all. On Thursday, the organizer needs to decide whether to cancel (and take advantage of the lower cancellation charges) or wait for Friday’s presumably more skillful forecast. If they wait, then on Friday they need to decide whether to cancel (and suffer higher cancellation charges) or go ahead and take the risk of the loss if the weather is bad.

In this example it is clear that rational decision-making requires not only an estimate of the probabilities of future outcomes, but also an understanding of how those estimates, and their skill, might change with subsequent forecasts. Similar examples can be constructed that relate to seasonal forecasts (e.g., a farmer having to decide whether to plant now or later) and climate forecasts (e.g., a government having to decide whether to build a flood defense now or later). Our example is idealized, and one could imagine factors that complicate the real-world decision-making situation. For instance, in reality, there may be forecasts available at greater frequency than daily, that allow a further option of cancellation late on Friday or early on Saturday, or the organizer may have the option to take out weather insurance to mitigate the loss if it occurs. In fact, for real-world decisions, it is seldom possible to write down every factor that influences the decision, let alone code them all into a mathematical framework, and practically all actual decisions are ultimately made using a subjective evaluation based on multiple inputs. As a result, this example should not be taken too literally. It nevertheless illustrates that the decide now or wait for the next forecast dilemma is an essential part of many decision-making situations related to weather and climate.

In section 2 we describe the single-stage cost–loss model in more detail, in the context of the example given above (that of an event organized for Saturday), within which the decision to make is whether to cancel the event. We then extend the model, following Murphy and Ye (1990) and Regnier and Harr (2006), and others, so that it can be used to address the decide now or wait question. General decision models, applying to many steps of forecast, with many possible actions and many possible weather outcomes, can become extremely complex, and the complexity may obscure the intuition behind the model and the solutions that it produces, potentially resulting in little useful insight and models that may never be used in practice. We will try to avoid this by making the minimum possible set of changes to the cost–loss model that allow us to answer the decide now or wait question, leading to a model that is simple and transparent enough that it can be readily understood. We will derive the model from basic considerations, so that it can be understood without first having to study the dynamic programming framework, and so that the derivation is intuitive for atmospheric scientists. This extended cost–loss model is then used to explore how to make a rational decision as to whether to decide based on the first forecast or wait for the second forecast, and to determine exactly what information is required for that decision.

In section 3, we consider the case in which the forecasts consist of normal distributions and are well calibrated, which allows certain simplifications in the modeling and leads to a straightforward implementation algorithm with which we can...
make the decide-now-or-wait decision. In section 4 we test the implementation algorithm from section 3 using a long series of synthetic weather forecast data. The synthetic data are created in such a way as to capture the relevant statistical structure of real temperature forecasts. In section 5 we perform similar tests on real forecast data and real observations using a rigorous out-of-sample method for testing the decisions. This is a much more challenging test, since the real forecasts and real observations undoubtedly do not perfectly fit the statistical assumptions on which the model is based. In section 6 we summarize the results and discuss the implications for weather and climate forecasting.

2. Cost–loss modeling

a. The single-stage cost–loss model

We now explain the single-stage cost–loss model so that we can introduce the concepts needed to extend it in section 2b below. The single-stage cost–loss model as used in atmospheric sciences assumes that a probabilistic forecast is available that gives the probability of two possible weather outcomes: $p$ for bad weather and $1 - p$ for good weather. The forecast probabilities are assumed to be well calibrated (i.e., we assume they have been adjusted based on what can be learnt from past performance of the forecast system) and so can be taken as the best estimate probabilities we have, and do not require further adjustment.

To analyze the model, one can consider the different possible outcomes as a function of the choices that could be made by the event organizer. Each outcome has a probability, based on the forecast, and a utility, based on the definition of the problem. The probabilities and the utilities can be combined to calculate the expected utility for each of the organizer’s possible choices, and the assumption in the model is that the organizer will opt for the choice with the higher expected utility. The utilities for each outcome are given in Table 1 and discussed below.

To apply the expected utility framework, first, we consider the choice in which the organizer goes ahead with the event. In this case there are two possible outcomes, depending on the weather, which are given different utilities in the model: good weather (probability $1 - p$) leads to no cost and no loss, and so is given a utility of zero, while bad weather (probability $p$) leads to a loss, and so is given a utility of $-L$, where $L$ is positive. The expected utility of going ahead with the event ($E_{\text{go, ahead}}$) is the sum of each probability multiplied by the corresponding utility, giving $E_{\text{go, ahead}} = (1 - p)(0) + (p)(-L) = -pL$.

Now we consider the choice in which the organizer cancels the event. In this case there are again two possible weather outcomes but this time both are given the same utility of $-C$, the cost of cancellation, where $C$ is positive. The expected utility for cancellation ($E_{\text{cancel}}$) is therefore $E_{\text{cancel}} = (1 - p)(-C) + (p)(-C) = -C$.

If the organizer seeks to maximize their expected utility, then the decision to cancel would be taken if the expected utility of canceling is greater than the expected utility of going ahead, $E_{\text{cancel}} > E_{\text{go, ahead}}$, which gives $-C > -pL$. Rearranging this expression leads to $p > C/L$. The conclusion is that, for the organizer to maximize their expected utility, they should cancel if the probability of bad weather is greater than a critical probability given by $p_{\text{crit}} = C/L$. This expression is the reason the model is referred to as the cost–loss model.

If $C$ is greater than $L$, then $p_{\text{crit}}$ is greater than 1, and the event will never be canceled because cancellation always has a lower utility than bad weather on the day. The interesting cases arise when $C < L$ and there is a trade-off between canceling and incurring the cost of cancellation, on the one hand, and not canceling and incurring the risk of bad weather and associated loss, on the other.

Models similar to the atmospheric sciences cost–loss model have been studied in other fields. For instance, economists have studied the question of whether to buy a stock today, given estimates of the probability of the stock price increasing or decreasing to certain levels tomorrow, and extensions of that question to multiple stages. See, for example, the discussion starting on p. 80 of Chambers and Lambert (2015), or the discussion starting on p. 95 of the textbook by Dixit and Pindyck (1994).

b. Extending the cost–loss model

We can extend the single-stage weather and climate cost–loss model, while staying within the expected utility framework, as follows. To make the explanation as readily understood as possible we will continue to use our illustrative example based on an event organized for Saturday. We now assume that two (lagged) weather forecasts are available for Saturday, one on Thursday and one on Friday. The utilities for each outcome are given in Table 2 and discussed below. The decision framework we now derive applies equally well to other types of forecast and other time periods, such as weather forecasts from Monday and Friday for Saturday, or climate forecasts for 2050 produced in 2030 and 2040.

On Friday, the organizer faces the same decision as is described in the single-stage cost–loss model: whether to cancel Saturday’s event. We will now write the utility of cancellation.
on Friday as $-C_1$, where the subscript 1 indicates cancellation 1 day in advance of the event, or, more generally, 1 forecast step in advance. The critical probability then becomes $p_{crit} = C_1/L$, and the organizer should cancel the event on Friday if the probability of bad weather exceeds $p_{crit}$ as before.

We now, additionally, move backward one step in time and imagine the organizer considering a weather forecast on Thursday, at which point they have the choice to either cancel there and then, or wait for Friday’s forecast. This is the decision that we will now analyze in detail. Canceling on Thursday leads to a cancellation utility of $-C_2$ and the interesting cases arise in this problem when cancellation on Thursday is cheaper than cancellation on Friday, which is in turn cheaper than last-minute cancellation on Saturday ($C_2 < C_1 < L$). Cancellation on Thursday being cheaper than cancellation on Friday ($C_2 < C_1$) leads to a dilemma for the organizer, particularly when the weather forecast on Thursday (for Saturday) is looking bad. There is now a trade-off for them between either canceling on Thursday and benefiting from Thursday’s cheaper cancellation fee or waiting for Friday to make a more informed decision.

From a mathematical point of view, the decision on Thursday is complex because it may be followed by, and needs to take account of the possibility of, having to make another decision on Friday, and that second decision would be based on information (Friday’s forecast for Saturday) that is not available on Thursday. To help make Thursday’s decision we must analyze how much we do already know on Thursday about what Friday’s forecast might be.

To analyze the trade-off involved in Thursday’s decision using expected utility, we first define four probabilities, $p_1$, $p_2$, $p’$, and $\hat{p}$. In the single-stage cost–loss model described above $p$ is used to represent the forecast probability of bad weather on Saturday, as evaluated on Friday. In the extended cost–loss model, we will now write the same probability as $p_1$ to indicate a 1-day forecast. We will also define the forecast probability of bad weather on Saturday, as evaluated on Thursday, as $p_2$.

From the point of view of Thursday $p_1$ is now not a single probability but is a random variable and has a range of possible probability values that are all the possible values that Friday’s forecast might take, given what we know on Thursday. For instance, if the forecast for Saturday, created on Thursday, is already saying that bad weather is very likely then $p_2$ will be known, and high, and we would already be able to predict that $p_1$ will most likely have high values, even though we would not know exactly the value it would take until Friday.

In this sense one could imagine creating a probabilistic forecast on Thursday for the range of values that $p_1$ might have on Friday, and indeed the simulation algorithm described in section 3 below, and applied in sections 4 and 5, involves making just such a probabilistic forecast of future forecasts. This probabilistic forecast captures how we think the probability of bad weather on Saturday will change from what we are predicting on Thursday, to what we might predict on Friday. From this probabilistic forecast for $p_1$, we will then evaluate the probability that $p_1$ will exceed the critical value $p_{crit}$, and we will call this new probability $p’$. Since exceeding the critical value leads to cancellation of the event on Friday, $p’$ is the probability that we would cancel the event on Friday, as assessed on Thursday.

In the single-stage cost–loss model, if the event organizer chooses to go ahead, because $p < p_{crit}$, then there is still the chance that the weather will turn out bad during the event. This happens with probability $p$ in that model. In the extended cost–loss model, we will again need to consider the chance that the organizer goes ahead but the weather turns out bad during the event, but we now need to evaluate it on Thursday so that it can form part of the basis for the decision to be made on Thursday. We will call this probability $\hat{p}$. From Thursday’s point of view, going ahead yet having bad weather, can arise from a range of values of $p_1$. For instance, we can imagine one case (on Friday) in which $p_1$ turns out only just below the threshold $p_{crit}$. In this case, the organizer would go ahead, but bad weather on Saturday is not that unlikely, since $p_1$ is still fairly high. On the other hand, we can imagine another case in which $p_1$ turns out far below the threshold $p_{crit}$, in which case bad weather on Saturday is more unlikely. $\hat{p}$ is the mean of the probability of bad weather, conditional on going ahead, over all such cases i.e., for different levels of $p_1$ in the range $[0, p_{crit})$. In summary, $\hat{p}$ is the probability, evaluated on Thursday, that if on Friday $p_1$ does not exceed $p_{crit}$, the weather on Saturday will nevertheless be bad.

Table 3 summarizes the definitions of $p_1$, $p_2$, $p’$ and $\hat{p}$ for reference. The meanings of $p’$ and $\hat{p}$ will become clearer in the context of the normal distribution example, discussed in section 3 below.

| Table 3. Definitions of the different probabilities used in the extended cost–loss model. |
|-----------------------------------------------|
| $p_{crit}$ | The critical probability in the single-stage cost–loss model, as used on Friday to decide whether to cancel Saturday’s event. |
| $p_1$ | The probability, evaluated on Friday, of bad weather on Saturday; when considered from the point of view of Friday, $p_1$ takes a single value; when considered from the point of view of Thursday, $p_1$ has a distribution of possible values. |
| $p_2$ | The probability, evaluated on Thursday, of bad weather on Saturday. |
| $p’$ | The probability, evaluated on Thursday, that on Friday $p_1$ will exceed $p_{crit}$. |
| $\hat{p}$ | The probability, evaluated on Thursday, that, if on Friday $p_1$ does not exceed $p_{crit}$, the weather on Saturday will nevertheless be bad. |

c. Expected utility analysis

Given the definitions of $p_1$, $p_2$, $p’$ and $\hat{p}$, we can now derive an expression for the expected utility of the two possible choices in the extended cost–loss model. The decision to be analyzed in this case is the decision taken on Thursday to cancel or wait for Friday’s forecast. The logic of the derivation is illustrated in Fig. 1.

First, we consider the choice of canceling on Thursday. This leads to a 100% chance of a utility of $-C_2$, and hence an expected utility of cancellation on Thursday ($E_{cancel\_Thursday}$) of $E_{cancel\_Thursday} = -C_2$. 
The first of these outcomes, canceling on Friday, occurs if, from the point of view of Thursday in order to feed into Thursday’s decision. Different probabilities, which must be evaluated from the point of view of Thursday, there are two outcomes: cancel on Friday, or decide to go ahead. These occur with different probabilities, which must be evaluated from the point of view of Thursday. Having waited to Friday, there are two outcomes: canceling on Friday, or decide to go ahead. These occur with different probabilities, different utilities, and different contributions to the expected utility. The expected utility for the decision to wait is the sum of these three contributions.

Second, we consider the choice of waiting for Friday’s forecast. Having waited to Friday, there are two outcomes: cancel on Friday, or decide to go ahead. These occur with different probabilities, which must be evaluated from the point of view of Thursday in order to feed into Thursday’s decision. The first of these outcomes, canceling on Friday, occurs if $p_1 > p_{\text{crit}}$, and incurs a utility of $-C_1$. From Thursday’s point of view the probability of $p_1 > p_{\text{crit}}$ occurring is $p'$ (by the definition of $p'$ given above), and so the contribution of canceling on Friday to the expected utility for waiting on Thursday is $-p'C_1$.

The second of these outcomes on Friday, deciding to go ahead, is more complicated since the utility is then affected by the weather outcome. Deciding to go ahead on Friday will only occur if $p_1 < p_{\text{crit}}$, which occurs with probability $1 - p'$. If the weather is good, the utility outcome is then zero, and the contribution to the expected utility is zero. If the weather is bad, which occurs with conditional probability $\bar{p}$ (by definition of $\bar{p}$ given above) then the utility outcome is $-L$. The contribution to the expected utility of waiting on Thursday from going ahead on Friday is therefore $- (1 - p')\bar{p}L$. The probabilities in this expression can also be understood using the definition of conditional probability, which states that $p(a \text{ AND } b) = p(a)p(b|a)$ and which we can apply here to say that the probability of going ahead AND having bad weather is equal to the probability of going ahead ($1 - p'$) multiplied by the probability of having bad weather, given that we have gone ahead ($\bar{p}$).

From the above considerations, the overall expected utility of waiting on Thursday is made up of three contributions from the three possible outcomes that waiting on Thursday may lead to. These are: canceling on Friday ($-p'C_1$), going ahead and having good weather (0) and going ahead and having bad weather $[-(1 - p')\bar{p}L]$. Combining the three contributions to the expected utility of waiting (and noting that one of them is zero) gives a total expected utility of waiting ($E_{\text{waiting}}$) of $E_{\text{waiting}} = -p'C_1 - (1 - p')\bar{p}L$.

We have derived expressions for the expected utility for both of the choices that present themselves on Thursday. We can now proceed to the final step in the analysis, which is to compare the expected utilities of the two choices. If the organizer seeks to maximize their expected utility, then the decision on Thursday to cancel would be taken if the expected utility of canceling is greater than the expected utility of waiting, in which case $E_{\text{cancel, Thursday}} > E_{\text{waiting}}$, implying

$$-C_2 > -p'C_1 - (1 - p')\bar{p}L.$$  
(1)

If we fully understand our forecasting system, the forecast skill, and how forecasts can change in time, then we can calculate $p'$ and $\bar{p}$, since they are just properties of the forecast. This inequality then determines whether to cancel, as a function of $L$, $C_1$ and the new parameter $C_2$. If we decide to wait, then come Friday the complexity of the decision on Thursday can be forgotten, and the decision on Friday can be made using the single-stage cost–loss model.

In summary, we have derived an expression that solves the extended cost–loss problem of whether to cancel on Thursday or wait for another forecast on Friday. It depends on the
calculation of two forecast quantities that are extensions of what is normally included in a probabilistic forecast. The first is $p'$, the probability (evaluated on Thursday) that the probability (evaluated on Friday) of bad weather (on Saturday) exceeds a critical threshold. The second is $\hat{p}$, the conditional probability (evaluated on Thursday) of bad weather (on Saturday), given that the probability (on Friday) of bad weather (on Saturday) does not exceed the critical threshold.

The probabilities $p'$ and $\hat{p}$ can be considered as properties of a probabilistic forecast and forecast system. They are both functions of two dimensions: these dimensions are the threshold level (of e.g., rainfall, temperature or wind) that defines bad weather, and the threshold probability from the single-stage cost–loss problem applied to Friday’s decision. In principle one could imagine routinely calculating numerical approximations to these two-dimensional functions every time a forecast is created. Values could then be read off to solve specific extended cost–loss problems as they arise.

In the next section we will consider the special case of forecasts consisting of normal distributions, in which the calculation of $p'$ and $\hat{p}$ becomes somewhat straightforward.

In the analysis above we considered the outcomes that follow Friday’s forecast, and then moved back one step in time and considered them from the point of view of Thursday. This stepping backward in time is the essence of the dynamic programming approach to solving multistage decision problems (Bellman 1957) and can be extended to solve problems that involve many time steps, many forecast outcomes, and many possible decisions, as described in detail in Murphy et al. (1985) and Wilks (1991) and applied in, for example, McLay (2008).

3. The normal distribution case

To apply the extended cost–loss model derived above in the simplest possible context, we will now consider a forecast system that produces forecasts consisting of normal distributions that are made on Thursday and Friday for Saturday. We will first discuss the statistical properties of these forecasts in some detail before presenting methods that can be used for applying the decision-making framework. Our method is similar to the Markov chain approaches described in e.g., Regnier and Harr (2006) and McLay (2008), but applies the Markov chain to the mean rather than the probabilities, following Jewson and Ziehmann (2004), which allows for considerable simplification.

a. Forecast properties

Since they consist of normal distributions, both Thursday’s and Friday’s forecasts can be described using a mean and a standard deviation: the mean represents the best single forecast, and the standard deviation represents the uncertainty around that forecast. For the forecast made on Thursday we write the mean and standard deviation as $m_1$ and $s_1$, and for the forecast made on Friday we write the mean and standard deviation as $m_2$ and $s_2$, where $s_1 < s_2$ since we assume Friday’s forecast is more accurate on average. We write the observation for Saturday as $a$ (for “actual”) and define the forecast errors as $e_2 = m_2 - a$ and $e_1 = m_1 - a$.

We will assume that the forecasts are well calibrated, by which we mean that they cannot easily be improved by further statistical postprocessing based on past forecasts and past forecast errors. This leads us to make three calibration assumptions about the statistical properties of the forecast. The first calibration assumption is that the means of the forecasts are unbiased, and so $E(a|m_1) = m_1$ and $E(a|m_2) = m_2$. Taking expectations of both these expressions and using the law of iterated expectations gives $E(a) = E(m_1)$ and $E(a) = E(m_2)$, from which we can deduce that $E(m_1 - m_2) = 0$ and $E(e_1 - e_2) = 0$.

The second calibration assumption is that the standard deviations of the forecasts match the standard deviations of the actual forecast errors.

The third calibration assumption is slightly more complex. To introduce it, we first define the change in the mean forecast from Thursday to Friday as $\delta = m_1 - m_2$. Using the assumptions given above, $E(\delta) = E(m_1 - m_2) = 0$. We also note that

$$\delta = m_1 - m_2 = (m_1 - a) - (m_2 - a) = e_1 - e_2.$$  

At the point in time that the decision is being made on Thursday, Thursday’s forecast, and hence $m_2$ and $s_2$, are known. The details of how $m_2$ and $s_2$ are created are not relevant, as long as they satisfy the assumptions given above. For instance, $s_2$ could have been estimated simply from analysis of past forecast errors or could have been derived from a statistical calibration scheme that merges information from past forecast errors with information from the ensemble spread (Jewson et al. 2004; Gneiting et al. 2005).

Friday’s forecast, however, will not be known on Thursday. We do nevertheless need to be able to estimate $s_1$ already on Thursday in order to estimate the variance of $\delta$, $V(\delta)$, since $V(\delta)$ is required for the algorithm described below. The simplest method for estimating $s_1$ on Thursday would be to use past forecast errors. Alternatively, one could investigate whether there might be information in Thursday’s ensemble spread to help predict $s_1$ (i.e., to predict the uncertainty around the next forecast, given the current ensemble spread), although this has never, to our knowledge, been explored. Another approach would be to estimate $V(\delta)$ directly from the ensemble spread: this approach has been considered in Jewson and Ziehmann (2004).

To derive an expression for $V(\delta)$ we will assume, as the third calibration assumption, that the forecast error $e_1$ must be independent of the change in the forecast $\delta$. The justification for this assumption is that if this were not the case then, on Friday, having observed the change from $m_2$ to $m_1$ (and hence the value of $\delta$) one would have information about $e_1$ that would allow one to improve the forecast $m_1$. We are assuming that any such improvements have already been made as part of the forecast calibration process, and hence that there is no longer any information about $e_1$ contained in $\delta$, and that $\delta$ and $e_1$ must be independent. Writing

$$e_2 = e_1 - \delta,$$

we can take variances of both sides. Since $e_1$ and $\delta$ are independent, by the argument above, there are no correlation terms on the RHS, giving:
\[ V(e_2) = V(e_1) + V(\delta) \]

and hence
\[ V(\delta) = V(e_2) - V(e_1) = s_2^2 - s_1^2, \]  

(2)

and in this way, we are now able to calculate the variance of the change \( \delta \) from the variances of the forecast errors. This variance is used in the algorithm described below.

A final set of assumptions we make, in addition to the assumptions related to good forecast calibration given above, is that the forecast errors \( e_1 \) and \( e_2 \), and the change in the forecast \( \delta \), are all normally distributed.

The assumption that \( e_1 \) and \( \delta \) are independent is analogous to the efficient market hypothesis in economics, which says that stock price changes are mostly unpredictable. More details are given Wilmott et al. (1995), Baxter and Rennie (1996), and Hull (2017).

b. Solutions for \( p' \) and \( \tilde{p} \)

We now describe how we can calculate \( p' \) and \( \tilde{p} \) given a forecast with the properties described above, which can then be used to make the cancel-or-wait decision. We present two methods: the first is based on simulation, and the second on numerical integration. The two methods have pros and cons. The simulation method is intuitively simple and forms a framework that could readily be extended to encompass different distributions and more complex decision problems, such as including distributions to capture the uncertainty in the parameters \( s_1 \) and \( s_2 \). The numerical integration method, on the other hand, is faster and more accurate, but would be more difficult to extend. For the problems we study below in sections 4 and 5 we have found that the simulation method is both fast enough and accurate enough, given standard computing resources. However, in the extension to multiple time steps simulation methods might be prohibitively slow and numerical integration might be preferable.

To define good and bad weather we will assume there is a given threshold value \( \theta \) of the forecast variable that separates bad weather from good weather, with values higher than \( \theta \) giving bad weather. An example would be temperature, where values above a given high threshold (i.e., a heatwave) may lead to the cancellation of an event.

1) Simulation Solution

Given the forecasts defined above we now describe a simulation algorithm that can be run on Thursday for calculating \( p' \) and \( \tilde{p} \) for this forecast system. The simulation algorithm estimates \( p' \) and \( \tilde{p} \) in a conceptually straightforward way by simulating many possible versions of Friday’s forecast for Saturday, given the information available on Thursday, and calculates \( p' \) and \( \tilde{p} \) from these many simulated forecasts.

We start by considering the forecast mean on Thursday, \( m_2 \), and in the first part of the simulation method we model how the forecast mean might change from Thursday to Friday, i.e., how \( m_2 \) changes to \( m_1 \). Since we know that \( E(m_1 - m_2) = 0 \), we have \( E(m_1) = E(m_2) = m_2 \) (this latter step because on Thursday \( m_2 \) is no longer random but is fixed by Thursday’s forecast) and we see that the distribution of possible values for \( m_1 \) will be centered around \( m_2 \). We also know \( V(\delta) \), the variance of the change in the forecast means, from Eq. (2) above, and we have assumed that \( \delta \) is normally distributed. As a result we can model the distribution of values that \( m_1 \) might take on Friday as a normal distribution with mean \( m_2 \) and variance \( V(\delta) \), which we write as \( N[m_2, V(\delta)] \). Each possible value of \( m_1 \) in this distribution corresponds to a possible probability forecast on Friday, consisting of a normal distribution centered around that value of \( m_1 \) with standard deviation \( s_1 \). We are modeling a distribution of possible distributions for Friday’s forecast.

This leads to an algorithm that can be used on Thursday for the calculation of \( p' \) and \( \tilde{p} \), as follows:

1) Derive \( s_1 \) and \( s_2 \), either from analysis of past forecast errors, or from the ensemble spread, or both. We will assume single values for both \( s_1 \) and \( s_2 \), rather than distributions, although the method could be generalized to deal with distributions of parameter uncertainty by simulating over different values for \( s_1 \) and \( s_2 \). Continue only if \( s_1 < s_2 \).
2) Calculate \( V(\delta) \) using \( V(\delta) = s_2^2 - s_1^2 \).
3) Given \( m_2 \), simulate \( Q \) values for \( m_1 \), using \( N[m_2, V(\delta)] \), where \( Q \) should be chosen large enough for good convergence of the results of the algorithm.
4) For each of the \( Q \) simulated values of \( m_1 \), use the corresponding forecast \( N(m_1, s_1^2) \) to calculate a value of \( p_1 \) (the probability of exceeding \( \theta \)), using the cumulative distribution function (CDF) for the normal distribution.
5) Count how many of the \( Q \) values of \( p_1 \) exceed \( p_{\text{crit}} \), to give \( R \).
6) Estimate \( p' \) as \( R/Q \).
7) For each of the \( Q - R \) forecasts for which \( p_1 \) does not exceed \( p_{\text{crit}} \), calculate the probability that the forecast variable exceeds \( \theta \), using the CDF for the normal distribution.
8) Estimate \( \tilde{p} \) as the mean of these \( Q - R \) probabilities.

In this normally distributed case, we see that \( p' \) and \( \tilde{p} \) are easily derived from modeling the distribution of possible future forecast distributions. This, in turn, is derived from an understanding of possible changes in the mean forecast, which in turn is derived from knowledge of the properties of the forecast errors.

The model used for the change in forecast means given above can be described mathematically as a single step of a random walk, or as a type of Markov process known as a martingale. Similar models can be used to model Brownian motion (in physics), or the logarithm of the changes in stock prices (in economics). A more detailed discussion of using random walks for both stock prices and expected weather outcomes is given in Jewson et al. (2005), and for stock prices in many standard finance textbooks such as Wilmott et al. (1995), Baxter and Rennie (1996), and Hull (2017).

2) Numerical Integration Solution

We now describe how \( p' \) and \( \tilde{p} \) can, alternatively, be calculated using analytical expressions and numerical integration, which would be faster and more accurate, but less easy to generalize. We start by deriving an expression for \( p' \).

On Friday, the probability of bad weather on Saturday is given by \( p_1 \). Given values for \( m_1 \) and \( s_1 \) (with \( s_1 \) assumed to be a
single value) and using the assumption that the forecast consists of a normal distribution, $p_1$ can be written using the cumulative distribution function of the standard normal distribution $\Phi$ as 1 minus the probability of good weather, giving:

$$p_1 = 1 - \Phi\left(\frac{\theta - m_1}{s_1}\right) = \Phi\left(\frac{m_1 - \theta}{s_1}\right).$$

If $p_1 > p_{\text{crit}}$ then the event will be canceled. The function $\Phi$ is monotonically increasing, and so this will occur for large values of $m_1$. Instead of using a threshold for $p_1$ we can therefore use a threshold for $m_1$ that we will call $m_{\text{crit}}$, defined by

$$p_{\text{crit}} = \Phi\left(\frac{m_{\text{crit}} - \theta}{s_1}\right).$$

Or the inverse:

$$m_{\text{crit}} = s_1 \Phi^{-1}(p_{\text{crit}}) + \theta.$$

The decision to cancel can now be based on $m_1 > m_{\text{crit}}$ instead of $p_1 > p_{\text{crit}}$.

The probability density of $m_1$ given $m_2$, as evaluated on Thursday, can be given in terms of the probability density of the standard normal distribution $\phi$, as

$$p(m_1|m_2) = \frac{1}{s} \phi\left(\frac{m_1 - m_2}{s}\right),$$

where we have defined $s = |V(\delta)|^{1/2} = \sqrt{V(\delta)}$. This is the distribution that we simulate from in step 3 of the simulation algorithm above.

The probability $p'$ that $m_1$ will exceed $m_{\text{crit}}$ is equal to 1 minus the probability that $m_1$ will be less than $m_{\text{crit}}$. The probability that $m_2$ will be less than $m_{\text{crit}}$ is given by the CDF of the distribution of $m_1$, evaluated at $m_1 = m_{\text{crit}}$, and so

$$p' = 1 - \Phi\left(\frac{m_{\text{crit}} - m_1}{s}\right) = \Phi\left(\frac{m_1 - m_{\text{crit}}}{s}\right).$$

This expression for $p'$ can be used to replace step 6 of the simulation algorithm.

We now derive an expression for $\hat{p}$. The probability $\hat{p}$ is the probability that the weather is bad, even if the event goes ahead, and we will write this as $P(\text{bad goes ahead})$. Using the law of total probability, this can be decomposed into the integral over all possible values of $m_1$ as

$$\hat{p} = P(\text{bad goes ahead}) = \int P(\text{bad} | m_1) p(m_1 | \text{goes ahead}) \, dm_1,$$

where $P$ is used to indicate a cumulative probability and $p$ is used to indicate a probability density. The first term inside the integral, $P(\text{bad} | m_1)$, is the cumulative probability of bad weather given $m_1$, and is given by the expression for $p_1$ given above. The second term inside the integral, $p(m_1 | \text{goes ahead})$, is the probability density of $m_1$, given that we go ahead. The probability density $p(m_1 | \text{goes ahead})$ is proportional to the probability density of $m_1$ given as $p(m_1|m_2)$ above in the range from minus infinity to $m_{\text{crit}}$, and is zero elsewhere. With renormalization of $[1/(1-p')]$ to ensure that it is a probability density, the probability density $p(m_1 | \text{goes ahead})$ is thus given by

$$p(m_1 | \text{goes ahead}) = \frac{1}{s(1-p')} \phi\left(\frac{m_1 - m_{\text{crit}}}{s}\right).$$

Putting these together in the integral gives

$$\hat{p} = \frac{1}{s(1-p')} \left[ \int \phi\left(\frac{m_1 - m_{\text{crit}}}{s}\right) \phi\left(\frac{m_2 - m_{\text{crit}}}{s}\right) \, dm_1 \right].$$

This expression can be evaluated using numerical integration as replacement for steps 7 and 8 in the simulation algorithm.

4. A synthetic forecast example

We now test the extended cost–loss decision algorithm using synthetic randomly generated forecasts and observations with appropriate statistical properties. There are two reasons for first using synthetic, rather than real, forecast data. First, by using synthetic data we can test the logic of the extended cost–loss decision framework and the normal distribution implementation algorithm, without also having to test whether any particular real forecast dataset fits the assumptions made in the derivation of the implementation algorithm. Second, by using synthetic forecasts we can derive results that are presumably as good as the results from the decision framework could ever be, because the data can be constructed so that the assumptions will be satisfied perfectly. These results can then be used as a benchmark against which to compare results from real forecast data. We use the simulation algorithm rather than the numerical integration solutions because we have the ambition to extend the model to allow for distributions for $s_1$ and $s_2$, and to forecast distributions other than normal.

a. Constructing the synthetic forecast dataset

The synthetic data need to satisfy various statistical properties in order to represent real forecasts sufficiently realistically. There are four conditions that the synthetic data need to meet: the synthetic forecasts and forecast errors need to be normally distributed, the forecasts need to be unbiased, the mean-square error (MSE) values have to be realistic in the sense that the 1-day forecast should on average be more accurate than the 2-day forecast, and finally $e_1$ and $\delta$ need to be uncorrelated, following the discussion in section 3a above.

To achieve these properties the synthetic forecasts and observations are created using the following steps, which work by first simulating Thursday’s forecast, then Friday’s forecast conditional on Thursday’s forecast, and then the observations conditional on Friday’s forecast. Simulating in this order makes it straightforward to create synthetic forecast data with the properties required, although other simulation methods could also be used such as a single step of simulation from a three-dimensional multivariate normal with appropriate means and covariances.

1) Assign values to $s_1$ and $s_2$, the forecast error standard deviations, with $s_1 < s_2$. 
2) Calculate $V(\delta)$ using $V(\delta) = \hat{s}_1^2 - \hat{s}_2^2$.
3) Simulate $D$ values for Thursday’s forecast mean $m_2$ using $N(0, \hat{s}_{2 \text{pop}}^2)$, where $\hat{s}_{2 \text{pop}}$ is the variance of values of $m_2$ and can be estimated from past forecasts.
4) For each value of $m_2$, simulate a corresponding value of Friday’s forecast mean $m_1$ using $m_1 = m_2 + \delta$ where $\delta$ is simulated using $N[0, V(\delta)]$.
5) For each value of $m_1$, simulate a corresponding value of the observation $a$ using $a = m_1 - e_1$ where $e_1$ is simulated using $N(0, \hat{s}_1^2)$.

That the synthetic forecasts generated in this way have the required statistical properties can be demonstrated as follows:

1) Friday’s forecast is unbiased because $E(m_1 - a) = E(e_1)$, and $e_1$ is simulated with mean zero.
2) Thursday’s forecast is unbiased because $E(m_2 - a) = E(m_1 - \delta - a) = E(e_1) - E(\delta)$, and both $e_1$ and $\delta$ are simulated with mean zero.
3) The $e_1$ and $\delta$ are uncorrelated because they are simulated from independent normal distributions.
4) The MSE of Friday’s forecast is $\hat{s}_1^2$ because $e_1$ is simulated with variance $\hat{s}_1^2$.
5) The MSE of Thursday’s forecast is $\hat{s}_2^2$ because $V(m_2 - a) = V(m_1 - \delta - a) = V(e_1) + V(\delta) = \hat{s}_2^2$.

We can also calculate the implied correlation between the two forecast means $m_1$ and $m_2$, which is given by $\left(1 + \frac{V(\delta)}{\hat{s}_{2 \text{pop}}^2}\right)^{-1/2}$.

We use the above algorithm to simulate $D = 10,000$ sets of two forecasts and one observation and we define “bad weather” to be temperatures over a threshold defined by the 70th percentile of the observed temperature distribution. To give a practical interpretation of this definition, one could imagine an event for which high temperatures on the day of the event might lead to immediate cancellation for health and safety reasons and would incur a loss in terms of refunds to paying participants. As a real-world example, the 2019 New York City triathlon, due to be held on 28 July 2019, was canceled at the last minute because of a prediction of a heat wave, and all participants were refunded their entry fees (CNN 2019).

We use values of $\hat{s}_{2 \text{pop}} = 3.011^\circ \text{C}$, $s_1 = 2.767^\circ \text{C}$, and $s_2 = 3.515^\circ \text{C}$, giving $V(\delta) = 4.697^2$ and a correlation between $m_1$ and $m_2$ of 0.582. These values are derived from the real forecasts used in the next section. The values could be very different for different forecast variables, and for different weather, seasonal and climate forecasts.

We compare the average utilities from applying the extended cost–loss model (which we label as extended) with that from three less sophisticated strategies, which are: always ignore Thursday’s forecast and wait until Friday before making a decision using the single-stage cost–loss model (which we label as always-fc1); always decide on Thursday using the single-stage cost–loss decision model and then ignore Friday’s forecast (which we label as always-fc2); and the more subtle strategy of using the single-stage cost–loss decision model on Thursday and then again on Friday if the event has not already been canceled (which we label as basic-twice). The basic-twice method is the most similar to the extended cost–loss decision model that we have derived, but neglects to take into account a proper analysis of the potential value of waiting for the next forecast when making Thursday’s decision. This is taken into account in the extended model.

b. Synthetic forecast results

Results for $C_1$ in the range from 0.1 to 0.8 are shown in the eight panels in Fig. 2. In each case $C_2$ is varied such that the ratio $C_1/C_2$ takes values from 1 to 1.6, while $L$ is always fixed at 1. Figure 3 shows the results for extended–basic-twice, along with bootstrapped confidence intervals. We see in Fig. 2 that extended gives the best results overall, as would be expected from the theory and derivations given in the previous sections, and because it brings more information to bear on the decision. For almost all parameter settings tested it gives higher average utilities than the simpler methods, and Fig. 3 shows that the differences between extended and basic-twice are significant at the 95% level in most cases, especially for lower values of $C_1$. In some cases, basic-twice gives slightly higher average utilities, but the differences are never significant. Basic-twice beating extended, but not in a significant way, can be explained by the two methods being very close for parameter settings that correspond to limiting cases, and that we have used a finite number of simulations. The relative ranking of the other methods varies with the parameter values: only extended is always in first or second place. This shows that extended does the best job of making use of the forecast data to make sensible decisions across a wide range of situations.

The results in Fig. 2 can be interpreted in more detail as follows. We start by considering the results for $C_1 = 0.1$ (the first panel in Fig. 2), which corresponds to canceling on Friday being cheap relative to the loss that might be incurred on Saturday: this makes canceling on Friday potentially attractive. When $C_1/C_2 > 1$ (the left-hand end of the horizontal axis) cancellation on Thursday is no cheaper than canceling on Friday, and so there is no reason not to wait to Friday. As a result, always waiting to Friday to make the decision (always-fc1) gives as good results as extended. The other two decision methods perform less well, because they use Thursday’s forecast, and this can only hinder making an optimal decision when canceling on Thursday is no cheaper than canceling on Friday.

For values of $C_1/C_2$ greater than 1 cancellation on Thursday becomes cheaper than cancellation on Friday, and the decision becomes a complicated one. All of the factors now come into play: the various costs, the skill of the forecasts, and the logic by which any decision made on Thursday needs to take into account how the forecast might change between Thursday and Friday (and what decision that might lead to on Friday). As a result, extended, which is the only method that takes all these factors into account, beats or matches the other three methods. The ranking of the simpler methods varies with $C_1/C_2$. As $C_1/C_2$ increases from 1 to 1.1 always-fc1 is soon overtaken by basic-twice and then by always-fc2. This is because always-fc1 ignores Thursday’s forecast, and this becomes increasingly unhelpful as canceling on Thursday becomes cheaper.

For the limiting case of large $C_1/C_2$, the results for extended, always-fc2, and basic-twice gradually converge, and only always-fc1 performs badly. This is because canceling on Thursday becomes very cheap, and the whole decision...
effectively becomes a question of whether to cancel on Thursday. Only methods that allow for that can do well.

Considering the other panels in Fig. 2: as $C_1$ increases, canceling on Friday becomes more expensive. Up to $C_1 = 0.4$ the results are qualitatively the same as they are for $C_1 = 0.1$. For $C_1 > 0.4$ the margin by which extended beats the other methods is reduced as canceling on Friday becomes more expensive and the value of making a good decision as to whether to wait until Friday is reduced.

For $C_1 > 0.6$ the results from the extended and basic-twice are very similar. This is because cancellation on Friday is so expensive that waiting to Friday to make the decision makes little sense: the decision can be made on Thursday. The subtle logic that extended uses to decide whether to wait until Friday becomes more or less irrelevant, and basic-twice actually does slightly better than extended in some cases (although not significantly so), presumably because of simulation noise.

Overall, we see from these results that extended always does well, but has the most impact in the situations in which all the options are potentially reasonable and the decisions on Thursday and Friday are both trade-offs. In the limiting cases in which canceling on Thursday or Friday is either cheap or expensive then the results from one or other of the simple methods are as good as extended. In a real situation, without detailed analysis, one would not know whether the parameters are in a limiting case and hence always using extended would make the most sense because it is the only method that works well in all cases.

5. A real forecast example

We now consider an example in which we apply the extended cost-loss decision framework derived in section 2, using the implementation algorithm described in section 3, to 30 years of real numerical model temperature reforecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF) and station observations for Stockholm, Sweden, from the Swedish Meteorological and Hydrological Institute (SMHI). We use the same set of decision problems and parameter ranges as was used to test synthetic forecasts in the previous section. Testing with real forecasts and real observations in this way is a much tougher challenge because of the uncertainties and unknown biases in real data, relative to our assumptions.

a. Constructing the real forecast dataset

Our forecast and validation datasets consist of 36-h and 108-h ERA-interim forecasts initialized at 1200 UTC (Dee et al. 2011) and related observed temperature values from 1979 to 2018. We only consider the 92-day period from 1 June to
31 August in each year. We use 1979–88 for calibration of the forecasts (giving 920 calibration cases) and 1989–2018 for out-of-sample validation (giving 2760 validation cases). We use forecasts with a wide spacing in lead time to magnify the impact of the method, so that the benefits can be seen more clearly above the noise in the results. The forecasts are calibrated using linear regression by regressing the observed values onto both the forecast and the previous forecast. The regression parameters are considered constant in time. We calibrate the mean of the ensemble while the standard deviation of the forecasts is derived from past forecast errors from the calibration period. This calibration method would be sufficient to ensure good calibration (as defined in section 3 above) if the forecast errors are genuinely normally distributed and homogeneous, and the calibration adjustments required are genuinely constant in time. In reality, the forecast errors are unlikely to be exactly normally distributed, may exhibit nonhomogeneity, and the ideal adjustments would likely not be constant time. Dependent on the level of misfit of these approximations, this might be expected to impact the effectiveness of the decision algorithm for this dataset. The calibrated forecasts consist of a mean and standard deviation for each of the two lead times: $m_1$, $s_1$ and $m_2$, $s_2$. The standard deviations of the forecasts, given by $s_1$ and $s_2$, are $s_2 = 3.515^\circ C$ and $s_1 = 2.767^\circ C$ From $s_1$ and $s_2$ we calculate the variance of $\delta$ as $\text{V}(\delta) = 4.697^\circ C^2$. We again define bad weather to be temperatures over a threshold defined as the 70th percentile of the observed data.

**b. Real forecast results**

We compare the average utilities from applying the extended cost–loss model over the 2760 validation cases with that from the three less sophisticated strategies as before. The results for all four methods are shown in Fig. 4, with Fig. 5 showing extended–basic-twice with 95% confidence limits. The results are similar to those shown for the synthetic data in Fig. 2. Overall, extended does better than the other methods and is the only method that always places best or second best. Extended again performs best, relative to the other methods, for $C_1 < 0.5$ and almost all the results are significantly better than basic-twice in this range (see Fig. 5). As for the synthetic data, extended does not give the highest utility score for every parameter setting. For example, for $C_1 > 0.4$ we can see that for some values of the ratio $C_1/C_2$ basic-twice does better than extended, but never significantly so. These are limiting cases in which the two methods give very similar results and basic-twice beating extended would seem to be due to noise in the data resulting from the finite sample size.

**6. Discussion and conclusions**

Probabilistic weather and climate forecasts can be used as input to decisions in various situations. The single-stage cost–loss model is an idealized form of a class of decisions that can be represented by an event organizer having to make a forecast-based decision by considering the trade-off between the cost of cancellation of an event one day in advance, and the risk of going ahead with the event and the weather turning out bad and causing a loss. Analogous situations, with the same logical structure, appear in many aspects of forecast-based decision-making, whether using weather forecasts, seasonal forecasts or climate projections. They also occur in many other branches of science, engineering, and economics.
The single-stage cost–loss model has previously been generalized in a number of ways, such as allowing for multiple possible actions (Murphy 1985), and multiple stages (Murphy et al. 1985; Epstein and Murphy 1987; Murphy and Ye 1990; Wilks 1991; Wilks et al. 1993; Wilks and Wolfe 1998; Regnier and Harr 2006; Roulin 2007; McLay 2011). In this study we have used a two-stage version of the model, involving two lagged forecasts, in which costs vary between the stages and there is a possibility of loss at the end of the second stage. We have considered the question of whether the forecast user should make their decision based on the first forecast or should postpone their decision until they see the second forecast. If, after consideration of the first forecast, a decision is made to wait, the second decision is then the same as that in the single-stage cost–loss model. Similar generalizations of the cost–loss model are considered by Murphy and Ye (1990) and Regnier and Harr (2006), and also arise in other fields.

We have analyzed the decide-or-wait decision that needs to be made based on the first forecast using expected utility. The process of using a forecast to make this decision requires the calculation of two new forecast quantities. One is \( p_0 \), the probability (evaluated using the first forecast) that the probability (evaluated using the second forecast) of bad weather will exceed a critical probability \( p_{crit} \), where \( p_{crit} \) is derived from the utilities of the different outcomes. The second new forecast quantity is \( \hat{p} \), the probability (evaluated using the first forecast) that, if (when we get the second forecast) we decide to go ahead with the event, the weather at the time of the event will nevertheless turn out bad.

These two quantities are nontrivial to calculate in general and require detailed modeling of the probabilities of the forecast system using Markov chains, as described in Regnier and Harr (2006) and McLay (2008). However, for the case in which forecasts and forecast changes consist of normal distributions, we have been able to derive a simpler method. In our method, the error statistics of the forecasts, which may already be known, can be used to derive the variance of forecast changes. The variance of forecast changes can in turn be used, when the first forecast becomes available, to run simulations of the possible distribution of probabilistic forecasts that will be produced in the second forecast. Based on this distribution of distributions, \( p' \) and \( \hat{p} \) can easily be calculated. Solutions for \( p' \) and \( \hat{p} \) based on numerical integration are also possible, which avoids the need for simulations.

We have tested our extended cost–loss decision algorithm on synthetic forecast data, and compared the utilities of the decisions that it makes with those made by three simple alternatives. The simple alternatives were 1) always decide using the second forecast, 2) always decide using the first forecast, and 3) decide using the first forecast, without taking into account that there will be a second forecast but then potentially reverse the decision using the second forecast (i.e., apply the single-stage cost–loss model twice on consecutive days). None of these simple methods correctly analyze the logic of whether to decide now or wait for the second forecast. The extended algorithm worked as expected and, overall, it gave better decisions than these simple alternatives. For some extreme parameter

![Fig. 4. Average utilities as in Fig. 2, but now based on real weather forecast data and observations for Stockholm.](https://example.com/fig4.png)
values it performed slightly less well than applying the single-stage cost–loss model twice on consecutive days (although not significantly so). We attribute this to the use of simulations in the testing and because in extreme parameter cases the subtle logic that the extended method uses becomes irrelevant, and simpler methods may be just as effective. The overall success of the extended method validates the underlying logic and also validates the effectiveness of the implementation algorithm for the case where the forecast errors are genuinely normally distributed and well calibrated. However, it is not surprising that the algorithm works well, since we are adding information to the decision process, and by using synthetic forecasts and observations everything except the randomness of the simulations is under our control.

We have also tested the method on real forecast data and real observations. This is a much more challenging test than using synthetic data, since real forecasts and observations will undoubtedly not follow our statistical assumptions precisely. To the best of our knowledge, this is the first time that the decisions from a multistage dynamic decision-making algorithm have been validated on real forecast data and real observations in this way. Once again, the extended cost–loss method outperformed simpler decision methods for most parameter values tested and performed best overall.

In summary, we have applied decision theory ideas and methods, specifically those related to dynamic programming, to the question of whether to decide now or wait for the next forecast, building on the work of multiple previous authors. In a situation in which waiting for the next forecast is an option, the probabilities in a probabilistic forecast are not enough to make a rational decision and need to be supplemented with additional information. In the case of well calibrated forecasts with normally distributed forecast errors, we have shown that the forecast MSE at the relevant lead times is sufficient to provide this information. The MSE can be used to derive the variance of the size of forecast changes, which in turn can be used to determine whether to wait for the next forecast. For nonnormal forecasts the picture is more complicated. For some nonnormal distributions there may be simplifying assumptions that can be used to derive simple algorithms as we have done for the normal distribution. In general, however, detailed nonparametric modeling of how forecast probabilities change from one lead time to the next is required. This general case has been considered by Regnier and Harr (2006) and McLay (2008).

One implication of our work, and prior work in this area, is that in order to realize the full potential of probabilistic forecasts for decision-making, forecast providers may need to consider providing additional forecast-change related information along with the forecasts that they supply. Forecast users could then use that information to make more rational decisions around the question of whether to wait for the next forecast. Those decisions might be made subjectively, with the additional information used as inputs, they might be made objectively using the extended cost–loss model we have described, or they might be made objectively using other decision theory models, as appropriate.

Acknowledgments. Author G. Messori was partly supported by the Swedish Research Council Vetenskapsrådet (Grant 2016-03724). We are very grateful to the various people, including the anonymous reviewers, who have made helpful suggestions that have greatly improved this work.
Data availability statement. The ECMWF forecast data can be obtained from ECMWF. The temperature observations for Stockholm can be obtained from the Swedish Meteorological and Hydrological Institute (SMHI).

REFERENCES

Baxter, M., and A. Rennie, 1996: Financial Calculus: An Introduction to Derivative Pricing. Cambridge University Press, 233 pp.

Bellman, R., 1957: Dynamic Programming. Dover Publications, 366 pp.

Buizza, R., 2001: Accuracy and economic value of categorical and probabilistic forecasts of discrete events. Mon. Wea. Rev., 129, 2329–2345, https://doi.org/10.1175/1520-0493(2001)129<2329:AAPEVO>2.0.CO;2.

Chambers, C., and N. Lambert, 2015: Introduction to dynamic belief elicitation. ACM SIGecom Exchanges, 14, 80–83, https://doi.org/10.1145/2845926.2845930.

CNN, 2019: The New York City Triathlon has been canceled because of the heat. CNN, https://edition.cnn.com/2019/07/19/us/heat-wave-new-york-triathlon-canceled-wxc/index.html.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: the data assimilation system. Quart. J. Roy. Meteor. Soc., 137, 553–597, https://doi.org/10.10102/qj.828.

Dixit, A., and R. Pindyck, 1994: Investment Under Uncertainty. Princeton University Press, 468 pp.

Epstein, E., and A. Murphy, 1987: Use and value of multiperiod forecasts in a dynamic model of the cost-loss ratio situation. Mon. Wea. Rev., 116, 746–761, https://doi.org/10.1175/1520-0493(1988)116<0746:UAVOMP>2.0.CO;2.

Gneiting, T., A. Raftery, A. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Mon. Wea. Rev., 133, 1098–1118, https://doi.org/10.1175/MWR2904.1.

Hagedorn, R., and L. Smith, 2009: Communicating the value of probabilistic forecasts with weather roulette. Meteor. Appl., 16, 143–155, https://doi.org/10.1002/met.92.

Hull, J., 2017: Options, Futures, and Other Derivatives. Pearson, 896 pp.

Jewson, S., and C. Ziehmann, 2004: Using ensemble forecasts to predict the size of forecast changes, with application to weather swap value at risk. Mon. Wea. Rev., 132, 1021–1031, https://doi.org/10.1175/1520-0493(2004)132<1021:UFETCP>2.0.CO;2.

Katz, R., 1993: Dynamic cost-loss ratio decision making model with an autocorrelated climate variable. J. Climate, 6, 151–160, https://doi.org/10.1175/1520-0442(1993)006<0151:DCLRD>2.0.CO;2.

Kernan, G., 1975: The cost-loss decision model and air pollution forecasting. J. Appl. Meteor., 14, 8–16, https://doi.org/10.1175/1520-0450(1975)014<0008:TCLCDM>2.0.CO;2.

McLay, J., 2008: Markov chain modeling of sequences of lagged NWP ensemble probability forecasts: An exploration of model properties and decision support applications. Mon. Wea. Rev., 136, 3655–3670, https://doi.org/10.1175/2008MWR2376.1.

Murphy, A., 1969: On expected-utility measures in cost-loss ratio decision situations. J. Appl. Meteor. Climatol., 8, 980–991, https://doi.org/10.1175/1520-0450(1969)008<0980:OEUMIC>2.0.CO;2.

—–, 1985: Decision making and the value of forecasts in a generalized model of the cost-loss ratio situation. Mon. Wea. Rev., 113, 362–369, https://doi.org/10.1175/1520-0493(1985)113<0362:DMATVO>2.0.CO;2.

—–, and Q. Ye, 1990: Optimal decision making and the value of information in a time-dependent version of the cost-loss ratio situation. Mon. Wea. Rev., 118, 939–949, https://doi.org/10.1175/1520-0493(1990)118<0939:ODMATV>2.0.CO;2.

—–, R. Katz, R. Winkler, and W. Hsu, 1985: Repetitive decision making and the value of forecasts in the cost-loss ratio situation: A dynamic model. Mon. Wea. Rev., 113, 801–813, https://doi.org/10.1175/1520-0493(1985)113<0801:RDMATV>2.0.CO;2.

Regnier, E., and P. Harr, 2006: A dynamic decision model applied to Hurricane landfall. Wea. Forecasting, 21, 764–780, https://doi.org/10.1175/WAF958.1.

Richardson, D., 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. Quart. J. Roy. Meteor. Soc., 127, 2473–2489, https://doi.org/10.1002/qj.49712757715.

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

Tena, E., and S. Gomez, 2011: Economic value of weather forecasting: The role of risk aversion. Top. (Madr.), 19, 130–149.

Terrado, M., and Coauthors, 2019: The weather roulette: A game to communicate the usefulness of probabilistic climate predictions. Bull. Amer. Meteor. Soc., 100, 1909–1921, https://doi.org/10.1175/BAMS-D-18-0214.1.

Ventura, J. A., 2019: Dynamic programming. Operations Research Methodologies, R. Ravindran, Ed., CRC Press, 168–194.

Wilks, D. S., 1991: Representing serial correlation of meteorological events and forecasts in dynamic decision-analytic models. Mon. Wea. Rev., 119, 1640–1662, https://doi.org/10.1175/1520-0493(1991)119<1640:RSCO>2.0.CO;2.

—–, 2014: Probability forecasting, quantitative decision modeling, and the economic value of forecasts. Edward S. Epstein Symp., Atlanta, GA, Amer. Meteor. Soc., 162, https://ams.confex.com/ams/94Annual/webprogram/Paper232199.html.

—–, and D. Wolfe, 1998: Optimal use and economic value of weather forecasts for lettuce irrigation in a humid climate. Agric. For. Meteor., 89, 115–129, https://doi.org/10.1016/S0168-1923(97)00066-X.

—–, R. E. Pitt, and G. W. Fick, 1993: Modeling optimal alfalfa harvest scheduling using short-range weather forecasts. Agric. Syst., 42, 277–305, https://doi.org/10.1016/0308-521X(93)90059-B.

Wilmott, P., S. Howison, and J. Dewynne, 1995: The Mathematics of Financial Derivatives: A Student Introduction. Cambridge University Press, 317 pp.