Norwegian
Meteorological
Institute

METinfo

No. 24/2022
ISSN 1894-759X
Meteorology

Verification Handbook

Fundamentals in meteorological verification

August 2022

Frank Thomas Tveter

The Norwegian Meteorological Institute

preface

When a meteorologist makes a forecast, and as the weather unfolds, it will soon enough become clear whether the forecast was useful or not. Checking if predictions agree with new observations is a form of *verification*, which is an intrinsic part of meteorology. But how does verification fit in with the process of developing meteorological knowledge? Are there alternative approaches available for developing knowledge that do not require verification? How do you make decisions that improve your weather forecast model? Should you include changes to the model even if they can not demonstrate any improvements? How do you know if your model has any prediction skill at all? Can you choose any verification score, or even create your own score? What happens if you develop and tune your meteorological model using one verification score, and later use another score when you compare it to other models? The purpose of this handbook is to focus attention on the fundamentals in the development of meteorological models using verification, and hint at some answers to these questions.

This handbook starts with a discussion on different approaches to accepting and improving knowledge. A general discussion on modeling and model prediction skill follows, with special focus on the relationship between model development and verification. A shortlist of meteorological verification scores, based on a comprehensive list maintained by the Joint Working Group on Forecast Verification Research, is presented at the end. The handbook also provides some instructive exercises. The answers provided are not absolute, but rather give an impression of how a meteorologist might respond to the questions. To encourage critical thought, one of the examples in this document reports a wrong decimal number.

Contents

Explanations

We use the term *explanation* to describe a statement that reveals a pattern in the real world, for instance “Thunder is caused by lightning”. An explanation is often a response to a question of the type “why did something happen?”, driven by the question “what will happen in the future?”. A *prediction* is a statement about the future inferred from an explanation, for instance if you observe a lightning strike, you could use your explanation to infer the prediction that a clap of thunder could follow shortly.

Let us assume that you have an explanation, which you can use to make predictions that can be observed. If the new observations are *in disagreement* with your predictions, your explanation is *falsified.* Otherwise they are by definition *in agreement* with your prediction and your explanation is *verified.* To *falsify* an explanation is to expose a disagreement between predictions and new observations, while to *verify* is to manifest an agreement. The process of making predictions and checking whether they agree with new observations is called *verification*, or sometimes *falsification.* The purpose of verification is always to falsify, and we use the term *validation* if the purpose is to verify. A *verification score* is a measure of how well new observations agree with predictions. *Prediction skill* is a verification score that can be related to some practical use. We say that an explanation *has* prediction skill if it yields good prediction skill scores.

An independent observation will usually not agree *exactly* with a prediction. Instead we associate each prediction with a probability distribution, and if the probability of making an observation is *unlikely* according to the distribution, we say that the observation *disagrees* with the prediction.

Your explanation is falsified if the joint probability of making the observations is *extremely unlikely* according to your predictions, otherwise it is verified.

A *correct* explanation can only be verified, and never falsified. A *valid* explanation *has* only been verified, and never falsified. An explanation that has been falsified, is *invalid.* A set of valid, consistent and unified explanations is called *knowledge.* A valid explanation that is consistent and unified with accepted knowledge is said to be *coherent.*

Exercise 1: The moon ring

John sees a ring around the moon and says that it will rain tomorrow.

(a) What is John's prediction?
(b) What is John's explanation?
(c) How may John verify his explanation?
(d) If John's explanation is verified, does that mean that it is correct?
(e) Can the explanation still be correct if it is falsified?
(f) Can John's explanation have prediction skill?

Answer 1: The moon ring

(a) John's prediction is that it will rain tomorrow.
(b) John did not provide any explanation.
(c) John could wait until tomorrow to see if it rains.
(d) No, you may never know if an explanation is correct.
(e) No, a correct explanation can never be falsified.
(f) Yes, it is of practical use to know if it will rain tomorrow.

Exercise 2: Black sheep

A tourist on a train in Scotland observes a black sheep. He has observed sheep in his homeland that are black or white.

Which of the following explanations are coherent?

(a) “Most sheep in Scotland are black.”
(b) “Cows look like black sheep in Scotland.”
(c) “All sheep in Scotland are white.”
(d) “Sheep in Scotland are black or white.”
(e) “Some sheep in Scotland are black on one side.”

Answer 2: Black sheep

(a) Coherent.
(b) Not coherent. Not unified.
(c) Not coherent. Falsified.
(d) Coherent.
(e) Not coherent. Not consistent.

Exercise 3: Public satisfaction

A score measures public satisfaction with a forecasting service.

(a) Is the score a verification score?
(b) Does the score measure prediction skill?

Answer 3: Public satisfaction

(a) No, the score depends on many other factors besides how well new observations agree with the predictions from the forecasting service.
(b) No, the score is not a verification score and it is not related to any practical use.

Likelihood

A joint probability of making a set of observations according to predictions, is called *likelihood*. Verification scores are usually designed to compare the underlying likelihood of different explanations.

We say that an outcome is *very unlikely* if the likelihood falls below $\sim 10\%$, and *extremely unlikely* if it falls below $\sim 1\%$.

In order to estimate the likelihood, it is necessary to know the probability distributions of the predictions. Statistics is a useful tool to estimate such probability distributions, if it is possible to make reasonable assumptions on random errors and error independence.

Law of large numbers

A simple statistical theorem which is often used in meteorology, is the *Law of large numbers*. The law of large numbers states that as the number of identically distributed, randomly generated variables increases, their sample mean (average) approaches their theoretical mean.

But note that even if the sample mean is defined, the theoretical mean does not always exist. If the theoretical mean does not exist, the sample mean will *not* approach a fixed value as the sample size increases, but rather keep varying as increasingly extreme values appear in the sample. We say that a distribution has large “wings” if the probability of encountering extreme values is significant.

Example 1: Undefined mean

In the following example, the theoretical mean, $E(x)$, is *undefined*, as the probability distribution, $p(x)$, has too large "wings",

$$\begin{aligned} p(x) &= \frac{1}{\pi\left(1+x^2\right)} \\ \int_{-\infty}^{\infty} p(x)dx &= \left.\frac{ArcTan(x)}{\pi}\right|_{-\infty}^{\infty} = 1 \\ E(x) = \int_{-\infty}^{\infty} x \cdot p(x)dx &= \left.\frac{Log(1+x_t^2)}{2\pi}\right|_{-\infty}^{\infty} = \text{undefined.} \end{aligned}$$

Meteorological systems may evolve in an non-linear manner, giving large "wings" in the derived probability distributions. It follows that in meteorology, you can never safely assume that there exists a theoretical mean.

Exercise 4: large "wings"

Assume that probability distributions with "wings" that fall faster than $\frac{1}{x^2}$ *as* $x \to \infty$*, have defined theoretical mean. Use a "small wing" criteria like* $p(100) \ll 0.0001$.

Which probability distributions have undefined theoretical mean?

(a) $p(x) \propto e^{-\sqrt{x^2}}$

(b) $p(x) \propto \frac{x^2}{1+x^4}$

(c) $p(x) \propto \frac{1}{\sqrt{1+x^2}}$

(d) $p(x) \propto x^4 e^{-x^2}$

Answer 4: Large "wings"

(a) Mean is defined.

(b) Mean is not defined.

(c) Probability function is invalid since it can not be normalized.

(d) Mean is defined.

Central limit theorem

The *central limit theorem* (CLT) establishes that, in situations when independent random variables with a theoretical mean μ and standard deviation σ are summed up, their normalized sum tends toward a Normal distribution even if the original variables themselves are not normally distributed. For example, if a meteorological observation is affected by a number of different independent error processes within the same magnitude, then the overall error distribution will have a *Normal distribution.* We may write the probability distribution of the sum, x, as

$$\begin{aligned} p(x) &= \frac{1}{\tilde{\sigma}\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\tilde{\mu})^2}{\tilde{\sigma}^2}} \\ \tilde{\mu} &= \mu \cdot N \\ \tilde{\sigma}^2 &= \sigma^2 \cdot N \end{aligned}$$

where N is the sample size. The highest probability density is located at the mean, $x = \tilde{\mu}$. Note that the central limit theorem does not apply if the theoretical mean of the independent random variables does not exist.

The probability of deviating from the mean of a Normal distribution by less than $1 \cdot \tilde{\sigma}$ is very likely (68%), by more than $2 \cdot \tilde{\sigma}$ is very unlikely (4%) while by more than $3 \cdot \tilde{\sigma}$ is extremely unlikely (0.3%).

Example 2: Binary events

For a binary (0 or 1) event with probability p, we have $\mu = p$ and $\sigma = \sqrt{p^2 + (1-p)^2}$. For example, if you repeatedly throw 32 coins on the table, the number of heads ($p = 0.5$) on each throw will follow an approximate Normal distribution with mean $\tilde{\mu} = 0.5 \cdot 32 = 16$ and standard deviation $\tilde{\sigma} = \sqrt{0.5} \cdot \sqrt{32} = 4$ Most times the sum will be between 12 and 20, and it is extremely unlikely that it is less than 2 or more than 30.

Note that if the sum is scaled by $\frac{1}{N}$, making it the average $\bar{x} = \frac{x}{N}$, we have that $\bar{\mu} = \mu$ and $\bar{\sigma} = \frac{\sigma}{\sqrt{N}}$. Taking the limit as $N \to \infty$ makes the variation approach zero, $\bar{\sigma} \to 0$, implying that $\bar{x} \to \bar{\mu} = \mu$, and in agreement with the law of large numbers.

Exercise 5: Observation system

An observation system is subject to many independent and small error sources with similar magnitude along the range of the instrument. Precipitation is known to add a large bias to observations.

How will the error distribution change if:

(a) The range is doubled?

(b) There is a 50% chance of precipitation?

Answer 5: Observation system

(a) $\tilde{\mu}$ increases by 100%, $\tilde{\sigma}$ increases by 40%.

(b) splits into two equal Normal distributions, one with a large bias.

Exercise 6: Event probabilities

An explanation predicts that there is a 20% chance of an event. Another explanation predicts a 80% probability. A sample of independent observations consistently indicates a 35% chance of the event. Assume Normal error distributions.

(a) Which sample size would have given a standard deviation of 15%?

(b) What is the likelihood for the 20% explanation in this case?

(c) What is the likelihood for the 80% explanation?

(d) Are the explanations falsified?

Answer 6: Event probabilities

(a) $N = 28$.

(b) The likelihood of the 20% explanation is 32%.

(c) The likelihood of the 80% explanation is 0.3%.

(d) The 80% explanation is falsified.

Exercise 7: Multiple case studies

A meteorologist performs 10 case studies comparing predictions from two explanations, and consistently observes a hit *(first explanation is best) in 70% of the cases. Assume that both explanations have the same quality. Use a Normal probability distribution for the number of hits.*

(a) What is the expected number of hits, $\tilde{\mu}$, and the standard deviation, $\tilde{\sigma}$, according to the assumption?
(b) Is the assumption very unlikely given the observations?
(c) How many cases are needed before the assumption is extremely unlikely?

Answer 7: Multiple case studies

(a) $\tilde{\mu} = 5$ and $\tilde{\sigma} = 1.6$.
(b) The assumption is *not* very unlikely.
(c) The assumption is extremely unlikely if there are 130 cases.

Exercise 8: Verification period

Consider an event where one explanation scores better than another over a verification period of one week. This independent event repeats itself in the next consecutive weeks. Assume that both explanations have the same quality. Use a Normal probability distribution for the number of events.

(a) What is μ and σ?
(b) When is the assumption very unlikely?
(c) When is the assumption extremely unlikely?

Answer 8: Verification period

(a) We have $\mu = 0.5$ and $\sigma = 0.7$.
(b) After 8 weeks.
(c) After 18 weeks.

Accepting explanations

A set of observations may match several patterns in the real world, and suggest several mutually exclusive explanations. The explanation that is *accepted* has a special status as it can be promoted without further support. There are many different approaches available for accepting explanations.

An approach for accepting explanations is defined by these methodological questions:

1. What is the purpose of the approach?
2. Which explanations can be considered?
3. Which explanation should be accepted?
4. When is the accepted explanation replaced?

The approach used to accept explanations in modern science is called *the scientific method*, and is based on *the legal principles*, while *the precautionary principle* is often used for environmental protection and *Murphy's law* is often used in engineering.

The scientific method

The scientific method is the principle method for the developing science.

The scientific method provides these methodological answers:

1. The purpose is to improve prediction skill.
2. Only consider explanations that can be falsified or verified.
3. The most coherent explanation is accepted.
4. The accepted explanation is not replaced until it is falsified.

The scientific method may appear conservative, in the sense that the accepted explanations gravitate towards knowledge already accepted by the scientific method. However, the primary focus of the scientific method is to falsify the accepted explanation, so that it can be replaced by an explanation with an even better prediction skill.

Exercise 9: The most likely explanation

New observations are 1% *likely according to the predictions from the accepted explanation, while they are* 99.9% *likely according to predictions from a new and spectacular explanation.*

(a) Is the accepted explanation falsified?
(b) Should it be replaced by the spectacular explanation?

Answer 9: The most likely explanation

(a) Yes, the accepted explanation is falsified.
(b) No, it should be replaced by the *most coherent* explanation.

The precautionary principle

The precautionary principle is suitable for developing explanations for complex systems like environmental protection, where it may be difficult to falsify the accepted explanation.

The precautionary principle provides these methodological answers:

1. The purpose is to remove threats.
2. Consider any explanation posing a threat.
3. The explanation posing the greatest threat is accepted.
4. The accepted explanation is never replaced.

Unlike *precaution*, the precautionary principle is used to remove threats that have never been verified before. Note that the precautionary principle does not require any verification, and the kind has a long history of promoting faulty explanations with dire consequences.

Exercise 10: Lead poisoning

Scientists suspect that Pb pellets are poisoning sea turtles. Sea turtles have low Pb levels, but it is feared that these will increase over time. Are the sea turtles being poisoned or not? Use the precautionary principle.

(a) Which explanation should be accepted?
(b) When is the accepted explanation replaced?

Answer 10: Lead poisoning

(a) The accepted explanation is that the turtles are being poisoned.
(b) The accepted explanation is never replaced.

Exercise 11: Stable gas discharge

There are many stable and harmless gases in the atmosphere. By combining known processes in the atmosphere, it can be argued that the release of a specific gas could cause an environmental disaster.

Is the release a threat to the environment, according to:
(a) the scientific method?
(b) the precautionary principle?

Answer 11: Stable gas discharge

(a) The release is not a threat according to the scientific method.
(b) The release is a threat according to the precautionary principle.

Murphy’s law

Murphy’s law states that “whatever can go wrong, will go wrong”.

Murphy's law provides these methodological answers:

1. The purpose is to prepare for any unfortunate outcome.
2. Only consider explanations that will be verified.
3. Accept all the explanations that may give unfortunate outcome.
4. Additional explanations can be accepted later.

Murphy's law is usually applied while engineering in complex environments, to prepare for unfortunate outcomes.

Exercise 12: The shark dive

A tourist books a dive with endangered sharks in a remote location. Which advice can be given based on:

(a) the scientific method?
(b) the precautionary principle?
(c) Murphy's law?

Answer 12: The shark dive

(a) You will be fine.
(b) Cancel the dive.
(c) Bring a knife, GPS and radio.

The legal principles

The *legal principles* are applied to specify laws and determine if an accused person is guilty of breaking the law. The scientific method, with it's many explanations, is in a sense a generalization of the legal principles where there are only two mutually exclusive explanations, namely that "the accused is guilty" or "the accused is innocent. Note that falsifying one explanation is equivalent to "proving" the other, by elimination.

The legal principles provide these methodological answers:

1. The purpose is to maintain support for the law.
2. Only make laws against crimes that can be verified.
3. Accept "accused is innocent", unless proven guilty.
4. The accepted explanation is reconsidered if proof changes.

Exercise 13: Littering

Which laws are acceptable?

(a) It is forbidden to litter the park.
(b) It is forbidden to bring litter to the park.
(c) It is forbidden to consider littering the park.

Answer 13: Littering

(a) Acceptable. The corresponding crime can be verified.
(b) Unacceptable. To "bring litter" can not be verified?
(c) Unacceptable. Intentions can not be verified.

Exercise 14: The 3 suspects

A crime has been committed and there are only 3 possible suspects. Suspect A has admitted the crime, but is a known liar. Suspect B may commit a terrible crime in the future. Evidence suggests suspect C.

Who is guilty according to:
(a) the scientific method?
(b) the precautionary principle?
(c) Murphy's law?
(d) the legal principles?

Answer 14: The 3 suspects

(a) Suspect C is guilty.
(b) Suspect B is guilty.
(c) All suspects are guilty.
(d) No suspects are guilty.

Meteorological models

Explanations can be used to predict properties of real life systems, for instance the temperature at a location. The representation of selected properties of a real life system is called a *model*[1]. A model uses explanations to formulate *processes* that change the properties over time, which are themselves properties of the model. The purpose of a model is to describe *previous*, *current* or *future* properties of the real life system. In meteorology, the purpose of a model is usually to predict the weather.

A numerical weather prediction model is typically formulated using millions of independent *model variables.* Each model variable has a specific influence on the model properties, and all the model variables must be determined before the model properties represent the real life system. The complete set of model variables is called the *model state.* The model state that corresponds to the state of the real life system, is called the *true state* or the *true values of the model variables.* The true state is always unknown, and can only be described using a probability distribution. An earlier estimate of the model state is called the *first guess* and sometimes the *background.* The difference between the new observations and the model equivalent based on an earlier estimate of the model variables, is called *the innovation.* The techniques used to determine the model variables, using a combination of the innovation and a first guess, is called *data assimilation.* The resulting model state is called *the analysis.*

A model may consist of other models. A simplified model of a process, is called a *scheme.* A surface scheme is for instance a combination of simplified surface process models. The process of developing a scheme is called *cooking.* Schemes are detailed using *parameters* that have to be adjusted. We say

[1] A *mock-up* is a representation of an imaginary system.

that parameters are *tuned* if they are adjusted during development, and *corrected* if they are adjusted as part of the weather forecasting operation. A scheme that relies heavily on parameters, is called a *parameterization.*

Exercise 15: Two point model

A model has two grid locations with temperatures 11.0 *and* 11.2 *degrees. A new temperature observation of* 10.0 *degrees is made at the center point.*

What is the value of:

(a) the model variables?
(b) the model equivalent to the observation?
(c) the innovation?

Answer 15: Two point model

(a) 11.0 and 11.2 degrees.
(b) 11.1 degrees.
(c) -1.1 degrees.

Model errors

A meteorological model does not contain all the information in the real life system, so the model representation can never exactly reconstruct the real life system. It follows that even if the true state matches the model state, there will still be a difference between the real life system properties and the properties reconstructed from the model. This difference is called the *representation error.*

Meteorological systems are non-linear and their current true state depends on any earlier true state. Some small changes in some variables may evolve to large changes later, while other large changes may decrease to small changes as time passes. These obscure dependencies on the initial state make non-linear systems unpredictable. The unpredictable property of non-linear systems was first discovered in meteorological models, and led to the development of *chaos theory*. Some physical processes are well described, while others are unknown or can not be accurately described using the model variables. Parameterizations use simplified linear processes to represent complicated non-linear processes in the real life system, making the model less chaotic than the real life system. The effects of processes can easily be over- or under-estimated in complex systems like the atmosphere. As all variables in a complex meteorological system have some effects on some other variables, a model may easily over- or under-estimated the effects that remotely associated model variables have on each other, for instance the effects of clouds on surface temperatures. The error contributions to the forecasted state, originating from the simplified modelling of physical processes, is called the *physical modeling error*.

Observations also have errors and so does the background, which together with the representation error and physical modeling error, contribute indirectly to the analysis error.

Unpredictable behavior is sometimes modeled by adding a random statistical error. However, a randomness assumption needs justification, for instance that underlying "error processes" are independent in time and place, which is usually not the case in meteorology. It follows from the obscure dependencies on all scales in meteorological systems, that any use of advanced statistics is prone to obscure errors. Advanced statistical assumptions and calculations should therefore be avoided in meteorology. Simple systematic errors in meteorological models are usually corrected using various schemes. There are for instance always systematic errors in the innovation. Wind observations at a location may for instance be systematically lower than the model equivalent, due to neighboring vegetation not accounted for in the model, and radiation measurement models will for instance not include all atmospheric processes that affect the observations. The innovation is typically corrected using data from the previous time period, for instance to correct the statistical dependencies on air mass and measurement geometry.

Exercise 16: The model errors

Assume that you have a numerical weather model.

Which errors are directly affected by:

(a) observation instrument degradation?

(b) increased analysis resolution?

(c) improved forecast model parameterization?

Answer 16: The model errors

(a) The observation errors.
(b) The representation- and background-errors.
(c) The physical modeling error.

Exercise 17: A global model

A simple meteorological model of the world has only one temperature value, T, which is 14 *degrees. The temperature in Oslo varies from* -10 *to* 20 *throughout a year and is observed to be* 10, 12 *and* 11 *degrees at noon on three consecutive days.*

(a) What is the representation error in Oslo?
(b) What is the observation error?
(c) What is the physical modeling error in Oslo?

Answer 17: A global model

(a) The representation error is about 20 degrees.
(b) The observation error is unknown.
(c) The physical modeling error is about 1 degree per day.

Model properties

Assume that the optimal value for one model variable has been estimated. This does not imply that the values of the other model variables also are optimal. It follows that if one property of the atmosphere is well described by the model, you can never also assume that another property automatically is well described too. Even if two properties are closely related by the same processes, you can not safely assume that quality somehow is transferred from one property to the other. If for instance a numerical weather prediction model is skilled at forecasting temperatures on the mountains, there is no guarantee that it will also be skilled at forecasting temperatures in the valleys. There could always exist some process that the model does not represent properly, which only affects one property and not the other.

You can never know if a model has any prediction skill at forecasting a specific property, before that property has been explicitly verified.

First principles operate on *accumulated* properties of a model. The motivation for using first principles appears to be that accumulated errors somehow cancel each other out on larger scales. A first principle is often a statement about some accumulated property combined with a simple scheme, for instance that the total amount of water in a complex system is constant. From a meteorological perspective, first principles are just another set of assumptions that must be verified before they can be used. Verifying first principles is extremely difficult, for instance how do you observe the total water content or how can you be sure that no processes modify the water content in your system? First principles are therefore never used in meteorology.

Exercise 18: Assumptions

A model is good at predicting the inland (2 meter) temperature.
Can you assume that the model will be good at predicting:
(a) skin temperature?
(b) cloud and precipitation?
(c) coastal temperature?
(d) maximum inland temperature?
(e) average inland temperature?

Answer 18: Assumptions

(a) No.
(b) No.
(c) No.
(d) No.
(e) No.

Model tuning

A meteorological model will typically use several schemes. Simple schemes often yield larger errors than the corresponding complex schemes, but they are easier to implement. A model therefore typically starts out with simple schemes, which are gradually improved as the model is developed. An important part of any model development is therefore to optimize the model parameters used by the schemes, for instance the parameters used in the parameterization of sub-model process schemes or in the bias correction schemes. When a parameter is tuned, it is usually adjusted until it results in the best verification score against a special dataset. The parameters are adjusted through trial and error, while trying to avoid *over fitting*. Adjustments that improve a verification score against a special set of observations are kept, while the others are discarded. Note that the special dataset verification score is no longer independent of the model, and a completely new and independent set of observations must be used when comparing it to other models.

Improving models

The prediction skill of the meteorological models used for operational weather forecasting, is improved using the scientific method. All experience shows that a large arbitrary change to a well tuned weather prediction model will make the model perform worse. The most coherent and accepted explanation is therefore that your changes to a model will make it perform worse. According to the scientific method, you have to falsify this accepted explanation before you can replace it by the explanation that your changes will improve the model.

> The only way you may claim that your changes to a model does not make it worse, is by demonstrating that your changes yield an improvement in the prediction skill.

According to the scientific method, the old model version can *only* be replaced by a new version that verifies better. New versions of a meteorological model usually have a change that is expected to improve the model. For instance, the increasing computational power makes it possible to better resolve physical processes in the model. The explanation that the new model version is best is therefore often the most "coherent" explanation. However, the scientific method explicitly states that the accepted explanation has to be falsified before it can be replaced by the most coherent explanation available. So again, even if a version with a better formulation is developed, it still has to be shown that the new version verifies better than the old version, before the corresponding explanation can be accepted.

Preserving prediction skill

It could be tempting to demonstrate that the new model version has "similar quality" to that of the old version, implying that you could choose either without affecting the prediction skill - and then select the new version of the model for some arbitrary reason. This approach, to *preserve* prediction skill, is in direct conflict with the scientific method, whose purpose is to *improve* prediction skill. But even if you ignore the difference in purpose, it is also an open question how you could falsify the explanation that the old model version is best, without showing that the new version is better. And is the explanation that two *different* models may have the *same* quality even coherent with chaos theory? It is well known that even the smallest changes can evolve to large changes in chaotic systems, which would yield different scores. The "similar quality" approach is therefore a completely different approach to developing knowledge.

The scientific method blocks you from replacing the accepted explanation until it is falsified, even if the new explanation has many other advantages and has demonstrated "similar quality". This stringency can appear as a weakness in the scientific method, but it is a strength. For instance, a meteorological forecasting system has many expensive components, and it can be difficult to show the benefit of each component individually. A "similar quality" criterion could easily be used to remove an expensive component, and thus "chip away" at the prediction skill of the system, in clear violation of the scientific method. Accepting the pragmatic attitude that the quality is the same for all practical purposes, is clearly not in agreement with the scientific method, and may result in a drift towards worse prediction skill. The correct decision would here be to *promote the accepted explanation*, namely that any such change would degrade the system - even if this explanation could not be demonstrated.

Climatology

The conceptual problem in numerical weather prediction, is to estimate the probability density function (PDF) for the true state of the atmosphere given the *available information*, and propagate this density function forward in time. The probability density function can be written as, $P(x_t|y \cap x_b)$, where x_t is the true state of the atmosphere, y is the observations and x_b is the first guess. At some point into the forecast, the initial state information, $y \cap x_b$, can no longer improve the forecast due to model limitations. This is the *deterministic lead time limit*. At this point, the density function should be written as $P(x_t)$, since conditioning on $y \cap x_b$ has no effect. This probability density function, or any property of it, is called the *climatology*. Although the climatology is independent of the initial state information, it may depend on other information, for instance location and time of year. An example of climatology is the mean temperature in Oslo in January. Every meteorological model has an intrinsic climatology, which either is specified directly or indirectly through model tuning.

> The model climatology is the probability distribution, or any property of it, that a forecast model will predict as the lead time passes the point where the initial state information is irellevant.

The model that predicts the best climatology beyond the deterministic lead time limit, will yield the best verification score for these lead times. The best climatology is the distribution that best matches the actual weather in the verification period. Model climatology can be estimated using the law of large numbers, by sampling over a sufficiently long time period, for instance over the last 40 years, and assuming that the future climatology will follow that of the past. However, as the weather development involves non-linear processes on all time scales, a theoretical climatology may not even exist and the sampled climatology will depend on the time period chosen. Estimating "climatological changes" on long time scales is outside the scope of this verification handbook, since that is an environmental protection problem which is usually solved using the precautionary principle, and therefore does not require any verification.

Exercise 19: The raining model

A model systematically predicts rain three days into the forecast.

(a) What could be wrong with the climatology?

Answer 19: The raining model

(a) The climatology could be too cold or dry.

Estimating model variables

In meteorology, the process of estimating the model variables using new observations and a forecast of the variables from an earlier cycle, is called *data assimilation*. The purpose of data assimilation is to estimate the analysis that gives the best model prediction skill.

The analysis is a representation of the probability density function for the *true* state of the atmosphere. The analysis is used as a starting point to predict the *future* state of the atmosphere.

All the information used to determine the model variables has an uncertainty. The main challenge in data assimilation is to estimate the properties of the combined probability distribution for the true values of the model variables given all the available information. For instance, assume that there are two independent sources of information, the first guess x_b

and the new observations y, each with their own error properties. There are several techniques available for estimating and analyzing the combined probability function, $P(x_t|y \cap x_b)$.

Example 3: Combined Normal probability density

If for instance you have a single observation and first guess characterized by an error with unbiased Normal distribution, we may write

$$\begin{aligned} P(x_t|y \cap x_b) &= N(\mu = \mu_a, \sigma = \sigma_a) \\ \mu_a &= x_b + H\sigma_b^2(H\sigma_b^2 H + \sigma_y^2)^{-1}(y - Hx_b) \\ \sigma_a^2 &= \frac{\sigma_y^2 \sigma_b^2}{\sigma_y^2 + H\sigma_b^2 H}. \end{aligned}$$

where H is the linearized forward operator which can be multiplied with the model state to calculate the model equivalent to the observation, σ_y is the observation error and σ_b is the background error. Observe that the new combined Normal probability density function has maximum value at $x = \mu_a$ and standard deviation σ_a.

Exercise 20: Highest combined probability

A simple model has one temperature variable with a first guess value of 10.1 *degrees and uncertainty* $\sigma_b = 2.5$ *degrees. An observation of* 12.5 *degrees is made with an uncertainty of* $\sigma_y = 0.5$ *degrees. Assume independent and unbiased Normal error distributions with unity forward operator* $H = 1$.

(a) What is the value with highest combined probability (μ_a)?
(b) What is the corresponding standard deviation (σ_a)?

Answer 20: Data assimilation

(a) 12.4 degrees.
(b) 0.49 degrees.

Deterministic models

In deterministic weather prediction, a single model state is estimated from the observations and first guess. The estimated analysis can be written as $x_a(y, x_b)$. The analysis is propagated forward in time by the model to form a forecast, written as $x_f(x_a, t)$.

Deterministic weather prediction models are developed to yield as good verification scores as possible. Another way of putting this is that the deterministic weather prediction models are developed so that they estimate the analysis that minimize the verification *risk*.

Verification risk

Let us assume that the forecast that yields the best verification score, is based on the analysis that yields the best verification score. Further assume that we have a dataset with combinations of observations, y, first guess, x_b, and the true state, x_t. If we select the N combinations with a specific set of values for the observations and first guess, we get a probability distribution of the corresponding true state, which can be written as $P(x_t|y \cap x_b)$. This implies that if we made new and accurate verification observations of all the model variables, these observations should also follow the probability distribution, $P(x_t|y \cap x_b)$. We can in principle estimate the optimal analysis that minimizes the verification risk, by using this probability distribution.

The verification risk is defined as the expected value of a *loss function*. The loss function defines the penalty associated with a possible outcome. The analysis, x_a, can be used directly in the loss function, since it is a function of the observations, y, and first guess, x_b. For instance if $x_a(y \cap x_b) = 10C$ and $x_t = 12C$, the quadratic loss function yields $l(x_a, x_t) = (x_a - x_t)^2 = 4C^2$. The verification risk for a given analysis, $B(x_a)$, can be approximated as the loss function averaged over the N combinations,

$$\begin{aligned} B(x_a) &= \int l(x_a, x_t) P(x_t | y \cap x_b) dx_t \\ &\sim \frac{1}{N} \sum_n l(x_a, x_t). \end{aligned}$$

Minimum verification risk is achieved when $B(x_a)$ is minimum, and that value of x_a depends on the definition of $l(x_a, x_t)$ and the distribution of x_t, $P(x_t|y \cap x_b)$. The analysis is developed and tuned so that it minimizes the overall verification risk. The overall risk, B, is the expected value of $B(x_a)$.

In deterministic weather prediction, the definition of loss is made by the developers of the forecasting system.

> The loss function of the verification score used in the development of a deterministic weather prediction model, will define the property of $p(x_t|y \cap x_b)$ that the analysis will be tuned towards.

Tuning a model using a verification score, will in other words make the model gravitate towards the optimal analysis for the underlying loss function of the verification score.

Mean squared error loss function

A common loss function used for continuous variables is the squared error. When applied to the analysis $l(x_a, x_t) = -(x_a - x_t)^2$, we have that

$$\begin{aligned} B(x_a) &= -\int (x_a - x_t)^2 P(x_t|y \cap x_b) dx_t \qquad (1) \\ &= -E((x_a - x_t)^s | y \cap x_b) \\ &= -x_a^2 + 2x_a E(x_t|y \cap x_b) - E(x_t^2|y \cap x_b) \end{aligned}$$

where we have introduced the expected function, E, for instance

$$E(x_t|y \cap x_b) = \int x_t P(x_t|y \cap x_b) dx_t.$$

The optimal analysis which minimizes the risk is then given by

$$\begin{aligned}\left(\frac{\partial B}{\partial x_a}\right)_{x_a} &= -2 \cdot x_a + 2 \cdot E(x_t|y \cap x_b) \equiv 0 \\ x_a &= E(x_t|y \cap x_b).\end{aligned}$$

The optimal analysis which minimizes the deterministic *mean squared error* loss is given by the *expected* state.

Mean absolute error loss function

Another common loss function is the absolute error, $l(x_a, x_t) = -|x_a - x_t|$. The mean error is then given by

$$\begin{aligned}B &= E(x_a - x_t|y \cap x_b)) \\ &= \int_{-\infty}^{x_a} (x_a - x_t) p(x_t|y \cap x_b) dx_t \\ &\quad + \int_{x_a}^{\infty} (x_t - x_a) p(x_t|y \cap x_b) dx_t\end{aligned}$$

The optimal deterministic analysis which minimizes the mean absolute error risk is given by

$$\begin{aligned}\frac{\partial B}{\partial x_a} &= \int_{-\infty}^{x_a} p(x_t|y \cap x_b) dx_t \\ &\quad - \int_{x_a}^{\infty} (x_t - x_a) p(x_t|y \cap x_b) dx_t = 0 \\ \int_{-\infty}^{x_a} p(x_t|y \cap x_b) dx_t &= \int_{x_a}^{\infty} p(x_t|y \cap x_b) dx_t.\end{aligned}$$

The optimal analysis is in other words the median state.

The optimal analysis which minimizes the deterministic *mean absolute error* loss is given by the *median* state.

Likelihood loss function

A common approach in deterministic data assimilation is to define the analysis as the state that maximizes the likelihood,

$$x_a \quad = \quad \text{arg max}_{x \in R}(P(x = x_t | y \cap x_b)).$$

The corresponding likelihood loss function can be written as,

$$l(x_a, x_t) \quad = \quad -log(P(x_t = x_a | y \cap x_b)).$$

Deterministic data assimilation seeks to maximize the analysis likelihood, which corresponds to minimizing the likelihood loss function.

Event based loss function

The end user may be interested in knowing which model best forecasts an event, for instance precipitation above a given threshold.

If the model equivalent and observation agree on the event, we have a *hit*. If the model equivalent did not predict an event we have a *miss*. A *false alarm* is when the model predicted an event that did not happen. A *correct negative* is when the model and observations had no event.

An intuitive score related to the probability of making a correct forecast, is for instance the *accuracy* or the probability that the event is forecasted correct. Let us assume unity forward operator. The event that the state is above the threshold, x_c, may be written as $x_a > x_c$. If we assume a loss function equal to -1 if we have a hit, and 0 otherwise, we get one of the two integrals,

$$\begin{aligned} B(x_a) &= -\int_{x_c}^{\infty} p(x_t | y \cap x_b) dx_t \text{ if } x_a > x_c \\ B(x_a) &= -\int_{-\infty}^{x_c} p(x_t | y \cap x_b) dx_t \text{ if } x_a < x_c. \end{aligned}$$

The lowest risk, $B(x_a)$, is achieved if x_a is on the side of x_c that is closest to the median of $p(x_t|y \cap x_b)$, suggesting that the optimal analysis is given by the median state.

The optimal analysis for event based loss functions, is suggested by the *median state.*

Competing loss functions

Tuning a model towards one verification score may in theory result in the worsening of another verification score. This could be a problem for very non-Normal probability density functions for the true state.

A deterministic numerical weather prediction system is tuned to score best with the verification techniques used to develop the system.

As most verification scores have the same optimal analysis when the probability density function for the true state is Normal, and since the Normal approximation is a common assumption, tuning towards one of these scores is not a major issue. Just note that if the “best model” for a specific property consistently depends on the verification technique, that could be an indicator of a non-Normal probability density function for the true state in the verification dataset.

Exercise 21: Optimal analysis

The probability density function of the true state given the available information, is a Normal distribution with a mean of 12.5 *degrees and a* 2.0 *degree standard deviation.*

What is the optimal:
(a) least squared error analysis?
(b) least absolute error analysis?
(c) maximum likelihood analysis?

Answer 21: Optimal analysis

(a) 12.5 degrees (expected).
(b) 12.5 degrees (median).
(c) 12.5 degrees (maximum).

Exercise 22: Camel analysis

The probability function of the true state is only defined at 3 values: 20% *at 10 degrees,* 30% *at 15 degrees and* 50% *at 20 degrees.*
What is the optimal:
(a) least squared error analysis?
(b) least absolute error analysis?
(c) maximum likelihood analysis?
(d) threshold analysis?

Answer 22: Camel analysis

(a) 16.5 degrees (mean).
(b) 15 degrees (median).
(c) 20 degrees (maximum).
(d) 15 degrees (median).

Exercise 23: Tuning

A sub-model process makes a temperature variable change from 10 *degrees to* 11 *degrees in* 30% *of the cases, and to* 12 *degrees in* 40% *of the cases. Assume that a single optimal parameterization value is related to the verification loss function in the same way as for the optimal analysis.*
Estimate an optimal deterministic parameterization value for the:
(a) squared error loss function?
(b) absolute error loss function?
(c) maximum likelihood loss function?

Answer 23: Tuning

(a) 11.1 degrees (expected value).
(b) 11.0 degrees (median value).
(c) 12.0 degrees (most probable value).

Probabilistic models

Probabilistic weather prediction estimates and predicts the probability function of the true state given the observations and first guess. Ultimately, the end user uses this probabilistic information in his own risk analysis. For instance, a painter may only want to paint if there is no chance of rain, while a commuter may drive by car instead of the bus if there is a large chance of rain. In risk analysis the user must define his loss function depending on the local consequences of the alternative decisions and outcomes.

In probabilistic weather prediction, the end user defines the loss.

Exercise 24: Decision theory

Opening a kiosk for a day costs $\$5,000$*, while keeping it closed costs* $\$1,000$*. If the kiosk opens and it does not rain it earns* $\$10,000$ *, if it rains it earns* $\$0$*. There is a* 10% *chance of rain.*

(a) How much will the kiosk make on average if it opens?
(b) How much will it make on average if it remains closed?
(c) Should the kiosk open?
(d) At what chance of rain does it loose money?
(e) At what chance of rain should it remain closed?

Answer 24: Decision theory

(a) $\$4,000$.
(b) $-\$1,000$.
(c) The kiosk should open.
(d) It looses money if more than 50% chance of rain.
(e) It should remain closed if more than 60% chance of rain.

Ensembles

A common approach to probabilistic weather prediction is to make a collection, or *ensemble*, of slightly different forecasts. Each forecast is called an *ensemble member* and indexed by the *ensemble number*, i, which runs from 1 to N, the total number of ensemble members. Each ensemble member forecast is made starting with a slightly different model state. The probability distribution of the true forecasted state is then represented using the distribution of the ensemble member forecasts.

Note that there is a fundamental difference between deterministic and ensemble models, namely that the deterministic model state represents properties of a probability distribution, usually the expected true state of the atmosphere, while each ensemble member represents a possible instance of the true state. While the deterministic model simply propagates the expected state forward in time, the ensemble model propagates a possible instance of the true state, and should therefore be subject to much more unpredictable behavior. It follows that if each ensemble member is verified individually against a verification technique that is for instances optimal for the expected state, each ensemble member will score worse than a deterministic model.

Ensemble mean

The ensemble mean of the predicted model equivalent, f_i, can be calculated using

$$\bar{f} = \frac{1}{N} \sum_{i=1}^{n} f_i.$$

As for the deterministic forecasts, the ensemble mean should be used when the verification technique is based on the squared error loss function.

Ensemble median

The ensemble median of the predicted model equivalent to each verification observation, is the center value which has just as many member values above as below. As for the deterministic forecasts, the ensemble median should be used when the verification technique is based on the mean absolute error or threshold loss functions.

Verification

Verification is often used synonymously with prediction skill in a meteorological context. The prediction skill should reflect how well the model predicts new observations, so that it can be used to identify the model that makes the best predictions in accordance with the scientific method.

A good verification score should for instance:

1. reflect the likelihood of making the observations,
2. be consistent over independent sets of observations,
3. worsen if we add random noise to model predictions,
4. be simple and robust without arbitrariness,
5. be calculated in the same way for all models being compared.

Ideally the verification score should be an *estimate* of the probability of having a set of new observations given the predictions. But this is difficult to calculate without making arbitrary (un-fair) assumptions, or without being sensitive to the weather situation.

Meteorological verification *compares* model predictions against the *same* observations, based on the *same* technique and using *fair* assumptions.

Discarding observations in a weather situation that your model was not “designed for”, is an example of an arbitrary decision which should not be used in a verification technique within the framework of the scientific method.

Imagine that two models have slightly different verification scores. One model could by random coincidence have a better score than another model. If random coincidences in the verification dataset dominate the verification score, it can not be used to identify the model that would make the best predictions in the future, thus defeating the purpose of the verification score within the framework of the scientific method.

> An improvement in verification score must be considered against the possibility that the improvement happened by chance.

For many complicated scores, the possibility that one model scores better than another by chance, is often judged subjectively based on long experience with that specific score. For this reason, the scientific method implicitly encourages a scepticism towards new verification scores.

Observations

As meteorological systems are non-linear, their error distributions often have large "wings". This is also the case for observation errors. For instance, instruments may freeze up and radiation measurements could be affected by cloud droplets. Some observations with extreme errors will eventually make it past the basic sanity checks. Verification scores that are sensitive to extreme errors can therefore be riddled with noise. A large dataset with independent observations would reduce noise, but could be impractical to accumulate. In order to make a verification score more conclusive, some sort of observation quality control is often applied to remove outliers.

It is easy to make obscure and arbitrary quality control decisions that favor a particular model, when processing observations for a verification score. The resulting score could not be used to identify the model with best prediction skill, and would therefore defeat the purpose of the scientific method. However, the score could still be useful in alternative approaches for accepting explanations, for instance to gain funding or support for a narrative. It follows that the observation processing must be *transparent*, so that it can be scrutinized by peers, to make sure it does not defeat the purpose of the scientific method. Any observation quality control criterion should be *symmetric* with respect to the models and the same observations should be used in calculating the verification score for all the models. For example, observations that deviated significantly ($> 3\sigma$) from *all* the models could be discarded and the observation locations should be defined by using predefined lists. Model information should typically not be used in the preprocessing of the observations.

The synoptic scale of a weather system is typically a thousand kilometers, and a weather situation can typically prevail for a week. Dangerous weather like storms are usually also the most difficult to predict, and therefore make large contributions on most verification scores. The verification of a limited area model in meteorology, will typically use data for at least a four week period, to catch several storms. As there can be seasonal variations in the intensity, efforts are also put into looking at seasonal variations, for instance using three months of data at a time, and comparing consecutive or inter-annual scores.

Exercise 25: Observation quality

Observations that deviate by more than 2σ from any one of the models are discarded.

(a) Is this observation quality control criteria symmetric?

Answer 25: Observation quality

(a) No. A model that has occasional large errors will be favored.

Exercise 26: Scatterometer winds

A scatterometer measures microwave backscatter from small waves, and estimates two opposite wind solutions. In a post-processing step, a global model forecast is used to select one of the solutions. The global model wind is then verified using the post-processed scatterometer observations.

(a) Would a comparison of verification score be fair to other models?
(b) Can the post-processed scatterometer observation be used to only verify other models?

Answer 26: Observation quality

(a) No.
(b) Yes, if they are all independent of the global model.

Persistence

Forecasting that the weather tomorrow will be the same as the weather today is called *persistence*. In meteorology, we say that a model does not have any prediction skill unless it out-performs persistence. The same criteria applies for trends, i.e. your model does not have any prediction skill if it does not clearly out-perform persistence of well known trends.

A model that does not consistently out-perform persistence has no prediction skill and should not be promoted.

Exercise 27: Prediction skill

Three models with significant prediction skills are verified. Model A scores 1.90*, model B scores* 1.80*, model C scores* 1.95*, persistence scores* 2.50*. The uncertainty in each score is* ± 0.1*.*

(a) Which of the models has best prediction skill?
(b) Which of the models has worst prediction skill?

Answer 27: Prediction skill

(a) Model B.
(b) Model A and C.

Underlying loss

Some verification scores can be associated with an underlying loss function that has a known optimal analysis. Tuning models using such verification scores will guide the model towards the associated optimal analysis.

It can be instructive to study the relationship between these verification scores, the underlying loss functions and the optimal analysis. This time we assume that the predicted model equivalent (forecast) is short so that the physical modeling error is negligible. The predicted model equivalent, f_i, will then correspond to the analysis, x_a. It is also reasonable to expect that the distribution of the verification observations, o_i, made of the model state, will follow $P(o_i = x_t|y \cap x_b) = N(\mu = f_i, \sigma)$.

Squared error verification

Squared error verification is a special case of a score with an underlying loss function with known optimal analysis, as it is also closely related to the falsification likelihood for continuous variables. If we assumed that the predicted model equivalent f_i had a Normal probability distribution with error σ, the probability density p_i that we made the observation o_i in the verification, could be written as

$$p_i \sim e^{-\frac{(o_i-f_i)^2}{2\sigma^2}}.$$

A measure for the combined probability of making a set of independent verification observations, is a strictly increasing function of the combined probability, and could for instance be

$$\begin{aligned} P &\sim \log\prod_{i=1}^{n}(p_i) \\ &\sim -\frac{1}{N}\sum_{i=1}^{n}(o_i - f_i)^2 \end{aligned}$$

where we note that σ just adds to the scaling of the probability and can be removed when we only want to compare different models.

Observe that the squared error verification score, P, corresponds to the (discrete) expected squared error loss, $B(x_a)$, when $o_i \rightarrow x_t$ and $f_i \rightarrow x_a$ (Eq. 1). The optimal analysis in this case is the expected state.

The squared error verification score is a measure for the probability of drawing the verification observations given that the predicted model equivalent had a Normal error.

The squared error verification score is the most popular verification score for continuous variables where we expect Normal error distributions.

If we verify the ensemble mean with the best mean squared error score, we may identify the ensemble model with the best ensemble mean. Note that this score is proportional to the probability of drawing the observations given the predictions, if we approximate the ensemble PDF using a Normal distribution centered around the ensemble mean. Verifying the ensemble mean using a mean squared error score is therefore a fair way of comparing likelihood of different models.

Mean absolute error verification

The mean absolute error verification score is the mean absolute error, $|o_i - f_i|$, given by

$$\begin{aligned} MAE &\sim \frac{1}{N}\sum_{i=1}^{N}(|o_i - f_i|) \\ &\sim E\left(E(x_a - x_t|y \cap x_b)\right)) \end{aligned}$$

which corresponds to the mean absolute loss function, when $o_i \rightarrow y$ and $f_i \rightarrow x_a$. The optimal analysis in this case is the median state.

Event based verification

Event based verification checks for instance values above or below a threshold. There is a hit if $f_i > x_c$ and $o_i > x_c$. Adding together the accuracy gives the event based loss function, when $o_i \rightarrow y$ and $f_i \rightarrow x_a$. The optimal analysis in this case is the median state.

Verification techniques

The purpose of the verification techniques within the framework of the scientific method, is to identify the model that has the best prediction skills. A verification technique compares the model predictions from different models, using the same observations. The purpose is to identify the model that makes predictions that gives the observations the highest likelihood. Some verification techniques are closely related to this likelihood, while other are better described as *sanity checks* used to identify strange model behavior. Some model properties are *continuous*, others are *event based* and some are multi-category. There are also some properties that are only available for probabilistic models.

In addition to objective verification scores, a meteorologist will also use other information, for instance *sanity checks* and *case studies.* A good tool for a sanity check is for instance scatter plots of observations vs forecasts. A sanity check has no practical use and is therefore not a skill, and it can often be manipulated so it is not used to compare models. In case studies, the meteorologist studies individual weather cases, typically badly forecasted weather. The purpose of case studies is to identify problems with the model that can be missed by the common verification techniques.

A list of the most common verification techniques used in operational weather forecasting is continuously updated by the *Joint Working Group on Forecast Verification Research.* This verification handbook contains a subset of these verification techniques at the time of writing, presented in a similar format for an easy reference. Note that some comments have been changed along with some terminology, so please consult the original source for the official comments and terminology of the working group and for references to the original publications.

Continuous properties

The verification techniques for continuous properties will be demonstrated on a sample dataset of 10 temperature observations, o_i, and their forecast equivalent, f_i, taken from Stanski et al. (1989),

Day	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
f_i (C)	5	10	9	15	22	13	17	17	19	23
o_i (C)	-1	8	12	13	18	10	16	19	23	24

Mean squared error

The mean squared error, MSE, measures the mean squared difference between the forecasts and observations. MSE, is defined by

$$\text{MSE} \;=\; \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2$$

where f_i is the forecasted model equivalent to the observation, o_i. The range is $[0, \infty\rangle$, a perfect score is MSE $= 0$ and the score unit is the square of the observation unit.

Example 4: MSE

For the sample data we have MSE $= 10C^2$.

If it is assumed that each forecast represents the mean of a Normal distribution with a specific standard deviation, the MSE will have a strictly decreasing dependence of the likelihood of making the observations given the forecasts. The MSE is therefore suitable for comparing models of the mean state within the framework of the scientific method. Used the other way around, the MSE is suitable for tuning meteorological models so that they match the mean state. If the mean squared error is used to tune a deterministic model, it will approach the mean state. If a probabilistic model is tuned using MSE, the mean of the probability distribution should be used in the calculation of the verification score.

Root mean square error

The root mean squared error, RMSE, estimates the average magnitude of the forecasting errors, and is defined by

$$\text{RMSE} \quad = \quad \sqrt{\text{MSE}}$$

The range is $[0, \infty\rangle$, a perfect score is RMSE $= 0$ and the score unit is the the same as the observation unit.

Example 5: RMSE

For the sample data we have RMSE $= 3.2C$.

The root mean square error is simple and familiar, since the score unit is the the same as the observation unit. It measures the "average" error, weighted according to the square of the error. It does not indicate the direction of the deviations. RMSE has a strictly growing dependence on MSE and therefore many of the same properties.

Mean absolute error

The mean absolute error, MAE, estimates the average magnitude of the forecast error, and is defined by

$$\text{MAE} \quad = \quad \frac{1}{N}\sum_{i=1}^{N} |f_i - o_i|$$

The range is $[0, \infty\rangle$, a perfect score is MAE $= 0$ and the score unit is the the same as the observation unit.

Example 6: MAE

For the sample data we have MAE $= 2.8C$.

The mean absolute error is simple and familiar. The MAE is optimal if the forecast represents the mean of the probability distribution of the true state given the available information. Use the other way around, the ME is suitable for tuning meteorological models so that they match the median state. If the mean absolute error is used to tune a deterministic model, it will approach the median state. If a probabilistic model is tuned using MAE, the median of the probability distribution should be used in the calculation of the verification score.

Mean error

The mean error, ME, or bias, estimates the average forecast error, and is defined by

$$\mathrm{ME} \quad = \quad \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)$$

The range is $[0, \infty\rangle$, a perfect score is $\mathrm{ME} = 0$ and the score unit is the the same as the observation unit.

Example 7: ME

For the sample data we have $\mathrm{ME} = 0.8C$.

The mean error is simple and familiar, also called the *additive bias*. It does not measure the magnitude of errors and does not measure how well the forecasts match the observations, i.e. it is possible to get a perfect score for a bad forecast.

Standard deviation of error

The standard deviation of error, SDE, estimates the standard deviation of the error deviation from mean, and is defined by

$$\mathrm{SDE} \quad = \quad \left(\frac{\sum_{i=1}^{N} f_i}{\sum_{i=1}^{N}(f_i - o_i - ME)^2}\right)^{\frac{1}{2}}$$

The range is $[0, \infty\rangle$, a perfect score is SDE $= 0$ and the score unit is the square of the observation unit.

Example 8: SDE

For the sample data we have SDE $= 3.2C$.

The standard deviation of error measures how well the forecasts match the observations, while correcting for any bias. The score is complementary to the mean error, and it is considered to be the more important and harder to improve.

Correlation coefficient

The correlation coefficient, COR, estimates how well the forecast values correlate with the observed values, and is defined by

$$\begin{aligned}\text{COR} &= \frac{\sum(f_i - \bar{f})(o_i - \bar{o})}{\sqrt{\sum(f_i - \bar{f})^2}\sqrt{\sum(o_i - \bar{o})^2}} \\ \bar{f} &= \frac{1}{N}\sum_{i=1}^{N} f_i \\ \bar{o} &= \frac{1}{N}\sum_{i=1}^{N} o_i\end{aligned}$$

The range is $[-1, 1]$, a perfect score is $r = 1$ and the score has no unit.

Example 9: COR

For the sample data we have COR $= 0.914$.

The correlation coefficient is a good measure of linear association or phase error. Visually, the correlation measures how close the points of a scatter plot are to a straight line. It does not take forecast bias into account, and it is possible for a forecast with large errors to still have a good correlation coefficient with the observations. The COR score is sensitive to outliers.

Multiplicative bias

The multiplicative bias, MB, estimates the average forecast bias scaled by the observation bias, and is defined by

$$\mathrm{MB} \quad = \quad \frac{\sum_{i=1}^{N} f_i}{\sum_{i=1}^{N} o_i}$$

The range is $\langle -\infty, \infty \rangle$, a perfect score is $\mathrm{MB} = 1$ and it has no unit.

Example 10: MB

For the sample data we have $\mathrm{MB} = 1.06C$.

The multiplicative bias is best suited for quantities that have 0 as upper or lower bound. It does not measure the magnitude of errors, and does not measure how the forecasts match the observations, i.e. it is possible to get a perfect score for a bad forecast.

Event based properties

A forecast may predict that an event will happen or not. Rain and fog predictions are common examples of such event forecasts. An event may also be a prediction above some threshold, for instance wind speed greater than 20 m/s at a specific location. Events do not have any unit.

To verify event based properties, we start with a contingency table.

		Observed		
		yes	no	total
Forecast	yes	hits	false alarms	forecast yes
	no	misses	correct negatives	forecast no
	total	observed yes	observed no	all

where *hits* are “events that were forecasted and that did occur” and denoted H, *misses* are “events that were not forecasted and did occur” and denoted

M, *false alarm* are “events that were forecasted but did not occur” denoted F, *correct negative* are “events that were not forecasted and did not occur” denoted C. The total numbers of observed and forecast occurrences and non-occurrences are called the marginal distributions.

The following contingency table with fabricated data is used to demonstrate the verification measures below.

		Observed		
		yes	no	total
Forecast	yes	82	38	120
	no	23	222	245
	total	105	260	365

Accuracy

The accuracy, ACC, or fraction of correct, estimates the fraction of forecasts that were correct, and is defined by

$$\text{ACC} = \frac{H + C}{H + C + F + M}$$

The range is $[0, 1]$, a perfect score is 1.

Example 11: ACC

For the sample data we have ACC $= 0.83$ indicating the 83% of the forecasts were correct.

Bias score

The bias score, BIAS, or frequency bias, estimates the forecast frequency of events compared to the observed frequency of events, and is defined by

$$\text{BIAS} = \frac{H + F}{H + M}$$

The range is $[0, \infty\rangle$, a perfect score is 1.

Example 12: BIAS

For the sample data we have BIAS = 1.14 indicating slightly over forecasting of rain frequency.

The bias score indicates whether the forecast system has a tendency to under forecast (BIAS¡1) or over forecast (BIAS¿1) events. It does not measure how well the forecast corresponds to the observations, only measures relative frequencies.

Probability of detection

The probability of detection, POD, or hit rate, estimates the fraction of the observed events that were correctly forecast, and is defined by

$$\mathrm{POD} = \frac{H}{H+M}$$

The range is $[0, 1]$, a perfect score is 1.

Example 13: POD

For the sample data we have POD = 0.78, indicating that roughly 3/4 of the observed rain events were correctly predicted.

The probability of detection is sensitive to hits, but ignores false alarms. It is very sensitive to the climatological frequency of the event, and good for rare events. The score can be manipulated by predicting more events to increase the number of hits. It should be used in together with the false alarm ratio (below). POD is also an important component of the Relative Operating Property (ROC) used widely for probabilistic forecasts.

False alarm ratio

The false alarm ration, FAR, estimates the fraction of the predicted events that actually did not occur, and is defined by

$$\mathrm{FAR} = \frac{F}{H+F}$$

The range is $[0, 1]$, a perfect score is 1.

Example 14: FAR

For the sample data we have FAR = 0.32, indicating that roughly 1/3 of the forecasted rain events were not observed.

Probability of false detection

The probability of false detection, POFD, or false alarm rate, estimates the fraction of events that were forecasted but not observed, and is defined as

$$\mathrm{POFD} = \frac{F}{H+F}$$

The range is $[0, 1]$, a perfect score is 1.

Example 15: POFD

For the sample data we have POFD = 0.15, indicating that roughly 15% of the forecasted rain events were incorrect.

Threat score

The threat score, TS, or critical success index, estimates wow well the forecast events correspond to the observed events, and is defined by

$$\mathrm{TS} = \frac{H}{H+M+F}$$

The range is $[0, 1]$, 0 indicates no skill and a perfect score is 1.

Example 16: TS

For the sample data we have TS = 0.575, indicating that more than half of the rain events where correctly forecast.

The threat score can be thought of as the accuracy when correct negatives have been removed from consideration, that is, TS is only concerned with forecasts that "count". The score is sensitive to actual hits, penalizes both misses and false alarms and it does not distinguish source of forecast error. The TS score depends on climatological frequency of events giving poorer scores for rarer events since some hits can occur purely due to random chance.

Equitable threat score

The equitable threat score, ETS, or Gilbert score, estimates how the forecast events correspond to the observed events accounting for hits due to chance, and is defined by

$$\begin{aligned} \text{ETS} &= \frac{H-R}{H+M+F-R} \\ \text{expected random hits} &= \frac{(H+M)(H+F)}{H+C+F+M} = R \end{aligned}$$

The range is $[-\frac{1}{3}, 1]$, 0 indicates no skill and a perfect score is 1.

Example 17: ETS

For the sample data we have ETS = 0.44. ETS gives a lower score than ET.

The equitable threat score adjusts for random chance which could be used to manipulate the ET score by forecasting more rain in a wet climate. It is often used in the verification of rainfall in NWP models because its "equitability" allows scores to be compared more fairly across different regimes. The ETS score is Sensitive to hits because it penalizes both misses and false alarms in the same way, it does not distinguish the source of forecast error.

Heidke skill score

The Heidke skill score, HSS, or Cohen's k, estimates the expected correct forecasts for random chance, and is defined as

$$\begin{aligned} \text{HSS} &= \frac{H + C - E}{H + C + F + M - E} \\ \text{expected correct} &= \frac{1}{N}\left[(H + M)(H + F) + (C + M)(C + F)\right] = E(2) \end{aligned}$$

The range is $\langle -\infty, 1]$, 0 indicates no skill and a perfect score is 1.

Example 18: HSS

For the sample data we have HSS $= 0.61$.

The Heidke skill score estimates the fraction of correct forecasts relative to is random chance. In meteorology, random chance is usually not the best forecast to compare to, it may be better to use persistence or some other standard.

Hanssen and Kuipers discriminant

The Hanssen and Kuipers discriminant, HK, or true skill statistic or Peirces's skill score, estimates how the forecast separate the events from no events, and is defined by

$$\text{HK} = POD - POFD$$

The range is $[-1, 1]$, 0 indicates no skill and a perfect score is 1.

Example 19: HK

For the sample data we have HK $= 0.63$.

The Hanssen and Kuipers discriminant uses all elements in contingency table. It does not depend on climatological event frequency. The Hanssen and Kuipers score can also be interpreted as (accuracy for events) + (accuracy for non-events) - 1. For rare events HK is unduly weighted toward the first term (same as POD), so this score may be more useful for more frequent events. It can be expressed in a form similar to the ETS except the hits random term is unbiased.

Odds ratio

The odds ratio, OR, estimates ratio of the odds of a forecasted event being correct, to it being wrong, and is defined by

$$\text{OR} \quad = \quad \frac{HC}{MF} = \frac{\frac{POD}{1-POD}}{\frac{POFD}{1-POFD}}$$

The range is $[0, \infty\rangle$, 1 indicates no skill and a perfect score is ∞.

Example 20: OR

For the sample data we have OR = 20.8, indicating that the odds of a predicted event being correct is over 20 times greater than the odds of a predicted event being incorrect.

The logarithm of OR is often used instead of the original value. The score takes prior probabilities into account, gives better scores for rarer events and is less sensitive to hedging. Do not use this score if any of the cells in the contingency table are equal to 0. The score is widely used in medicine but not common in meteorology.

Odds ratio skill score

The odds ratio skill score, ORSS, or Yule's Q, estimates the improvement of the forecast over random chance, and is defined by

$$\text{ORSS} \quad = \quad \frac{HC - MF}{HC + MF}$$

The range is $[-1, 1]$, 0 indicates no skill and a perfect score is 1.

Example 21: ORSS

For the sample data we have ORSS $= 0.91$.

The odds ration skill score is independent of the marginal distribution, so it is difficult to hedge.

Rare event based properties

A rare event is characterized by a long time between each event.

Deterministic limit

The deterministic limit, DL, estimates the length of time into the forecast in which the forecast is more likely to be correct than incorrect. The deterministic limit is defined, for categorical forecasts of a pre-defined rare meteorological event, to simply be the point ahead of issue time at which, across the population, the number of misses plus false alarms equals the number of hits, i.e. threat score or critical success index =0.5. The base rate (or event frequency) should also be disclosed. Re calibration of the forecast is often necessary for useful deterministic limit measures to be realized.

As they provide a clear measure of capability, deterministic limit values for various parameters may in due course be used as year-on-year performance indicators, and also to provide succinct guidelines for warning service provision. They could also be used as the cut-off point to switch from deterministic to probabilistic guidance.

Extreme dependency score

The extreme dependency score, EDS, estimates the association between forecast and observed rare events, and is defined by

$$\text{EDS} \;=\; \frac{2\log\left(\frac{H+M}{H+C+F+M}\right)}{\frac{H}{H+C+F+M}} - 1$$

The range is $[-1, 1]$, 0 indicates no skill and a perfect score is 1.

The extreme dependency score converges to 2η-1 as event frequency approaches 0, where η is a parameter describing how fast the hit rate converges to zero for rarer events. The EDS is independent of bias, so should be presented together with the frequency bias.

Multi-category properties

Multi-category properties are events that have several discrete values, for instance light rain and heavy rain. We denote the category count, N, by two indexes, one for the forecasted category and one for the observed category. For instance n_{ij} is the number of cases where the forecast predicted category i while the observed category was j. Further we define n_{i*} as the total number of forecast cases with category i while n_{*i} as the total number of observed cases with category i. We may write $n_{i*} = \sum_{j=1} N n_{ij}$ and $n_{*j} = \sum_{i=1} N n_{ij}$.

The event property is a special case of a multi-category property, with two categories, event or no event. In this case, the number of hits, H, is given by n_{11}, false alarms, F, is $n_{1,2}$, misses, M, is n_{21} and correct negatives,C, is n_{22}. The continuous property is another special case, with an infinite number of categories.

Histogram

The histogram shows the probability distribution of the forecasts compared to the observations. This sanity check shows similarity between location, spread, and skewness of forecast and observed distributions, but does not give information on the correspondence between the forecasts and observations. Histograms give information similar to box plots.

The histogram is simple and intuitive, although it can be misleading since it is heavily influenced by the most common category.

Multi-category accuracy

The Multi-category accuracy, MACC, estimates the accuracy of the forecast in predicting the correct category, relative to that of random chance, and is defined as

$$\mathrm{MACC} \;=\; \frac{1}{N}\sum_{i=1}^{n} n_{ii}$$

The range is $[0, 1]$, a perfect score is 1.

The score is simple and intuitive, but can be misleading since it is heavily influenced by the most common category.

Multi-category Heidke skill score

The Multi-category Heidke skill score, MHSS, estimates the accuracy of the forecast in predicting the correct category, relative to that of random chance, and is defined as

$$\mathrm{MHSS} \;=\; \frac{\frac{1}{N}\sum_{i=1}^{N} n_{ii} - \frac{1}{N^2}\sum_{i=1}^{N} n_{i*}n_{*i}}{1 - \frac{1}{N^2}\sum_{i=1}^{N} n_{i*}n_{*i}}$$

The range is $\langle -\infty, 1]$, 0 indicates no skill and a perfect score is 1.

The multi-category Heidke skill score estimates the fraction of correct forecasts relative to random chance. The score requires a large sample size to make sure that the elements of the contingency table are all adequately sampled. In meteorology, at least, random chance is usually not the best forecast to compare to, it may be better to use persistence or some other standard.

Multi-category Hanssen and Kuipers discriminant

The multi-category Hanssen and Kuipers discriminant, MHK, or true skill statistic or Peirces's skill score, estimates how the forecast separate the events from no events, and is defined by

$$\mathrm{MHK} \quad = \quad \frac{\frac{1}{N}\sum_{i=1}^{N} n_{ii} - \frac{1}{N^2}\sum_{i=1}^{N} n_{i*}n_{*i}}{1 - \frac{1}{N^2}\sum_{i=1}^{N} n_{*i}^2}$$

The range is $[-1, 1]$, 0 indicates no skill and a perfect score is 1.

The score is similar to the Heidke skill score (above), except that in the denominator the fraction of correct forecasts due to random chance is for an unbiased forecast.

Probable event based properties

A probable event forecast gives a probability of an event occurring, p_i, with a value between 0 and 1, where i is the observation index.

Brier score

The Brier score, BS, estimates the magnitude of the probability forecast errors, and is defined by

$$\mathrm{BS} \quad = \quad \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2.$$

The range is $[0, 1]$, a perfect score is 0.

The Brier score measures the mean squared probability error. It is sensitive to the climatological frequency of the event, i.e. the more rare an event, the easier it is to get a good BS without having any real skill.

Relative operating property

The relative operating property, ROC, visualizes the ability of the forecast to discriminate between events and no events. The probability forecast is divided into bins using probability thresholds, for instance 0.05, 0.15, 0.25. The POD is then plotted against the POFD for each bin, and the area under the ROC curve is frequently used as the ROC score. The range is $[0, 1]$, where 0.5 indicates no skill and a perfect score is 1. The perfect ROC curve travels from bottom left to top left of diagram, then across to top right of diagram. The diagonal line indicates no skill. The relative operating property is not sensitive to bias in the forecast.

Ranked probability score

The ranked probability score, RPS, estimates how well the probability forecast predicted the category that the observation fell into, and is defined as

$$\mathrm{RPS} \quad = \quad \frac{1}{M-1}\sum_{m=1}^{M}\left[\left(\sum_{k=1}^{m} p_k\right) - \left(\sum_{k=1}^{m} o_k\right)\right]^2$$

where M is the number of forecast categories, p_k is the predicted probability in forecast category k, and o_k is an indicator (0=no, 1=yes) for the observation in category k. The range is $[0, 1]$, a perfect score is 0.

The ranked probability score penalizes forecasts that are more severe when their probabilities are further from the actual outcome. When there are only two forecast categories, the RPS is the same as the Brier Score.

Ensemble probability distribution properties

Ensemble models that forecast a probability distribution may for instance have their median and mean verified using techniques for continuous variables. It may also be interesting to look at a sanity check for the ensemble distribution.

Rank histogram diagram

The rank histogram diagram, RHD, or Talagrand diagram, estimates how well the ensemble spread of the forecast represent the spread of the observations. The rank histogram diagram is constructed using the following procedure:

1. Rank the N ensemble members from lowest to highest.
2. Identify which bin the observation falls into at each point.
3. Tally over many observations to create a histogram of rank.

The idea is that each ensemble member represent an equally likely instance, so the observations should spread equally between the members.

The rank histogram diagram can be interpreted using the following table:

Shape	**Interpretation**
Flat	ensemble represents forecast uncertainty
U-shaped	ensemble spread is too small
Dome-shaped	ensemble spread is too large
Asymmetric	ensemble contains bias