Decoding the intermittent atmosphere: surprise removal and information flow

JOHN R. LAWSON*

Cooperative for Mesoscale Meteorological Studies, University of Oklahoma, OK, USA

NOAA/OAR/National Severe Storms Laboratory, OK, USA

Valparaiso University, Indiana, USA

COREY K. POTVIN

Cooperative for Mesoscale Meteorological Studies, University of Oklahoma, OK, USA

NOAA/OAR/National Severe Storms Laboratory, OK, USA

KENRIC NELSON

(GOES HERE)

Submitted July 19, 2022

*Corresponding author address: Dept., Institution, Address, City, State/Country.

E-mail: john@jrl.ac

Abstract

The commonly used Brier Score is inappropriate for rare events, including similar probabilistic scores that use mean-squared error (such as the Ensemble Fractions Skill Score- CHECK). We present information theory as the correct scoring rule, particularly important for rare or common events, due to the logarithmic score's desired properties of additivity, propriety, and (?). The decomposition of a satisfactory measure of skill (cross-entropy score) similar to that in the Brier Score is the combination of information gained by discrimination (sharpness of the probability forecast), negated by error from reliability (introduced by issuing a non-binary forecast probability) and uncertainty (a combination of observation, model, and inherent uncertainty). This is analogous to the machine-learning "bias–variance" tradeoff, where overfitting occurs when discrimination skill is more than negated by reliability error (and vice versa for underfitting).

To demonstrate the importance of using the correct score, especially in a regime of rare events, we use Lorenz's simple 1963 model of chaos in an intermittent ("bursty") regime to generate simulated forecasts and observations. When evaluated with both Brier and information-theory methods, we find increasingly diverging evaluation for each forecast dataset as the event becomes rarer. We answer the question of how to pick the appropriate scoring rule by showing how information theory fits our intuition better, and forms a framework in which information flow (uncertainty or ignorance removal) can be evaluated from forecast to decision-maker to the end-user. The authors' aim is not to introduce yet another skill score to the meteorological community, but rather review existing frameworks and explain in plain language why the Brier Score is the wrong score for rare events.

1. Introduction

A wooden fence under wind load has two main states: in the a priori state of stable, or knocked over. There are only two states (just about) and the system can only move one way forward in time and the state cannot reverse (or we don't care if it does), and cannot recover without outside input (entropy, organisation, etc). The axiom of choice is everything at any minute: a decision tree that becomes more and more complex. Information entropy in bits is identical to the number of decision-tree splits. It takes time to navigate each split (time to process a more uncertain forecast). These are "on the knife-edge" scenarios where entropy is maximised (0.5) for self-entropy (i.e. for one event). How does this link to physical information? Think of the atmosphere at time zero encoding a message with minimum of (category, timescale, space-scale) of e.g., tornado. The scales are always there implicitly, but are better explicit as is typical. This is physically manifested in our conceptual model of an eddy growing, moisture building, self-organisation (emergence, entropy, etc). As chaos grows rapidly for e.g. tornado predictability or the path of the storm (inner-predictability or self-entropy/self-information), the parent storm may become stable (Supercell). The watersheds converge to stable or not stable, but the sub-system of the tornado gives little information of its formation/it is unpredictable. How does it communicate (Weak Links?) between layers of our conceptual model? An observer at the ultimate end point (centre of time/space radii) will have a log/exp curve of knowing whether to shelter or not simply by staring out the window. They could adjust that with climo (it's May, or it's snowing), though that increases the surprise (lower average entropy, higher self-entropy). Best is gathering information upstream. Hence we have myriad sources of information (satellites) to initialise NWP models that guide forecaster

to issue a polygon that let local decision makers decide: 1 or 0, shelter or not? The journey of the message (symbolically) is noisy. This is why chaos theory demands we use ensembles. We get erroneous information. If too much uncertainty is communicated (reliability or calibration error/information loss), it erodes from the information gain from overlapping with the true ”pdf” (i.e., 70% chance of tornado; RES, resolution, or DSC discrimination). The error that can't be removed, or the *a priori* information entropy, is UNC or uncertainty. This is used to determine a baseline for a skill score, but keeping the raw bits allows a schematic of an information river with various dams that reduce the water level at points along the route. We want to optimise the system in the long run so DSC is maximised, REL is minimised, UNC is also minimised (but outside the scope of the system unless a new baseline UNC can be created with pre-processing). We don't know the true UNC so that's the difference between IGN and XES: the uncertainty term (can this always be broken down? Key maths question, maybe for Kenric). But less uncertainty or more certain forecasts closer to 0/1, and/or removing uncertainty earlier in the river so the downstream dams can be faster or more confident, means less redunancy, higher information rate, less choice so less time,. But then surprise is larger, related to cost-loss ratio where surprise can be quantified with bits but also financially. This dictates the probability at which to act. How do we optimise this message? If it were 50pc each time, that's the worst. If it were 0.1pc, let's say, that's climatology and psychologically stupid. Surprise and catastrophic error in coding (Dulles v Dallas), (english v mandarin examples?). We want to optimise information flow in the noisy channel. How should we ”send the information” (can't change the weather encoding the message, so that bit is more like Cybernetics). This is conceptually used for verification (how well can we decode the message?). Optimisation for both sending and receiving in the channel hence is more social

science/communication. How many ensemble members to create a good resolution of the pdf/uncertainty (information dimension? Vertices in info space? Quantisation of 0.0-1.0 probs?), also kernel dressing. (This avoids the curse of dimensionality, where members need to go to the powers (?) of the dx/dy/dz/dt increase.) This can be evaluated ”purely” with comparing models and baselines, or via cost-loss ratios in units of bits or monetary currency. Decomposing the DSC/REL is important to remove UNC surprise to be dealt with separately. In AI, there is the bias-variance relationship with over- and under-fitting. Is this the same? Too tight to the learning set is overcertainty or confidence, or trying to maximise DSC too much. Too lax from the set is more agile and protects against surprise, but is underconfident in trying to minimise REL fraction of total surprise XES. Do we optimise this in the same way? Is it possible to use AI/ML and information gain (random forest uses decision trees and info gain, others?) to choose the best number of members, given finite resources. Also balance with grid spacing? Address dx versus members with this framework. Finally, can we observe the subprocesses, such as tornado formation, such as highlighting where information transfer is highest and where observations in the field would help the most. Could be where Lyapunov exponents are largest, or air parcels' orbits diverge rapidly (bifurcations in flow). This is the axiom of choice, or the tipping potted plant.

Numerous papers already exist (lit review in next section). These are verification etc. This paper starts with a demonstration that info theory is the correct scoring framework for forecasts. We use the Lorenz-63 toy-model of convection in an intermittent (”bursty”) mode to show critical deviation in forecast evaluation between the Brier Score and Cross-Entropy Score. We then fix these findings into a conceptual model of information flow and a framework of optimising detection of the signal (a hazard we want to predict, say) above

the noise (error in model, observational, initial- and lateral-boundary conditions). Once we reduce forecasts to the prediction of a hazard (message) from an ”alphabet” of weather states, there are many analogues between information theory – as applied to fields such as circuitry and communications – and the evaluation and optimisation of a predictive system. The largest hurdle is jargon; hence, we provide analogues between terms used in a variety of fields. The authors hope this encourages more interdisciplinary understanding given the connection in mathematics.

The aim is not to introduce a new score, but rather reiterate many previous papers and demonstrate the utility applied to the field of severe weather. However, the findings apply to any forecasting system, particularly relevant to those predicting the state of/within a chaotic, intermittent system such a tornado development.

Motivation. Evaluating weather models is an important way to determine which development NWP to operationalise. However, evaluators are tasked with concluding skill as objectively as possible, whilst inexoriably introducing subjectivity through their choice of scoring rule. Many scores can be mislead (ETS is not equitable;). A scoring rule must meet three criteria (why? See Benedetti): propriety, (additiveness?)... A **proper** score cannot be hedged: as in, a forecaster split between issuing a forecast of rain or no-rain over a number of forecast times is punished for choosing a probability of 50% each time. Only issuing the best-guess probability is correct. An **additive** score ...(unitless). A **what** score is WHAT? Hence, BS is wrong. XES is right (unitless, additive, etc). Indeed, following Benedetti [cite], we show Brier Score is a second-order approximation of the logarithmic score at the foundation of the XES, and speculate BS's longevity stems from its lower computational demand than scores that use a logarithm. (Harold Brooks, pers. comm.)

2. BACKGROUND

Conversion of terminology - see table

Amongst the many traditional methods of evaluating weather models (REF review paper), the Brier Score (Brier 1950 REF) is commonly used (REFS to show this) to evaluate probabilistic forecasts. The score can be divided into components, for instance, into X and Y (alternative derivation). The focus of this paper is the decomposition into reliability (REL), discrimination (DSC) [1]. The components eluciadate facets of the forecast that are blurred by the summation of these facets. Further, we show how these components have very practical and immediate use in the fields of machine learning and uncertainty communication.

a. Reliability

This is also called (Table XX) calibration (REF, e.g. Weijs), variance (? - machine learning), (more? radar? info theory?). This is a punishment for a probability error, such as issuing a forecast of 10% for all instances of rain that ultimately occurred 20% of that forecast period.

b. Discrimination

Also called resolution, mutual information (info teory), bias (ML ?)... It is the only component measuring a positive score for the forecast's performance.

[1] Herein, we use discrimination (DSC) rather than resolution (RES) that is often used elsewhere. This is due to the ambiguoity with horizontal grid-resolution, often termed resolition colloqually

c. Uncertainty

This is also called (Table XX) entropy, uncertainty, background noise (info theory) and irreducible error in machine learning. There are many papers that have recognised the superior suitability of information-theoretical scoring rules over commonly used (but sub-optimal) scores.

d. Information Theory

These components are insightful and recognisable in the meteorological community. The authors are not introducing a bespoke new score in the present manuscript, but rather identifying a better analogue already used in other fields and well documented in older meteorological literature. We hope to motivate this theoretical section.

Everything that occurs can be reduced to the **axiom of choice**: it is (represented by 1), or it is not (a 0). The choice of binary digits (*bits*) is represented by an unbiased coin with probability $P = 0.5$ of landing of heads or tails. How much uncertainty is there about which outcome will occur? How surprised will the observer be upon observing the result? These can be quantified (Shannon REF) using *information entropy*,

$$H = -\log_2 P \tag{1}$$

where the use of base 2 in the logarithm determines the units of H as bits. Other literature may use log or ln interchangably. Conversion between information units can be done via (EQN). Let's say we drop the coin causing it damage, and start to predict which side the coin lands without any (information) about how the coin may now be biased. Table XX shows the observer's predictions (made before all tosses as a straight guess *en masse*) compared to the result. Let's interpret the components. First, uncertainty (UNC)

is equivalent to Eqn. (REF) if we use the base rate of observations as the frequency of occurrence, (o bar). In information theory, this is the entropy of the observation dataset. For forecasts, a higher value (reaching 1 bit) represents the most uncertain forecast as in a coin flip (probability of 0.5) over a long time series. This is because, on average, picking heads each time will surprise the observed as often as it will not. A biased coin will mean one guess (heads or tails, depending on how the coin is biased) yields less surprise as a result. However, as the event becomes rarer – let's say the coin is heavily biased such that tails only occurs 1% of the time – while the average entropy (UNC) is low (if we know the frequency of heads before issuing a forecast, it's best to guess heads by default), we are very surprised when heads does turn up. Accordingly, Eqn. XX yields a much higher number via $H = -log_2 0.01 = 6.64 maybebits$ despite the lower series entropy of $H = \sum_{t=1}^{T} -log_2 0.01$, giving XYZ.

3. Brier Score as an Approximation

Show DS — BS.

4. Fractions Skill Score as an Approximation

Replace FBS with log? Or at least discuss how FSS is plagued with the same issues. We propose a 'fractional ignorance' skill score (poster citation) where the scoring rule can be swapped out.

5. Application

Why is this immedaitely applicable?

In the forecast process – say, whether to shelter from a tornado – each stage (NWS, EMs, public, decision-makers) acts on incoming evidence to reduce uncertainty. Humans do not begin entirely ignorant of the atmosphere's evolution: some states cannot occur (e.g., snow in 25 deg C), and analogues from past events can give a general sense of the flow regime (but due to sensitivity to initial conditions, analogues are nothing more than initial guides to constrain uncertainty somewhat). Discussed more esoterically later as conceptual model.

6. Methods

Maths of Brier and IG (XES) and not allowing 0.0/1.0 probs. The windowing is like FSS to show how the forecasts must have a temporal and spatial definition to make sense. (Citation, or just makes logical sense)

a. Lorenz-63

The Lorenz-63 (REF) system is a toy-model of convection in which Lorenz found chaos (Gleick, Lorenz?). When ρ ? is set to values (OF WHAT), the system is in an intermittent state. Hence, events such as "parameter Z exceeding its 90th percentile" become more difficult to predict (UNC) due to heightened sensitivity to initial conditions. This provides a source of simulated NWP models for our purposes.

$$X = o \tag{2}$$

$$Y = f \tag{3}$$

$$Z = p \tag{4}$$

$$\tag{5}$$

We run 500 simulations, with various modifications to create experiment with a variable each:

- Before running each simulation, we stochastically modified the X parameter at initialisation, X_0, such that each simulation has a different value between x and X.
- Ensembles of various sizes were created by randomly subsampling from the simulations.
- To emulate the temporal and spatial windowing in eFSS (eq or REF), we employ a moving window of varying (?) size. Quantisation.
- Drift (model error - did we use it?)
- Tuning parameter. The intermittency was changed by varying between X and Y. IMPORTANT - can we convert this to the change in base-rate frequency? Maybe determine the "climate" beforehand. Need to create longer climate and show why we chose that long (representative). (rename experiments as rare, common etc)
- (What else?) List a table of experiments?

We combine across these variables in the next section. This is a blunt tool. Can we combine the results in another way? Need to add degree of obs uncertainty too.

7. Results

Here, we present...

a. Preliminary work

Showing the quantisation and initial demo of how the L63 model works

b. Comparison of BS versus XES

Compare the plots (BS, BS REL/DSC, XES, XES REL/DSC, BSS, XES)

8. Synthesis: decoding the atmosphere

How does this fit into the conceptual model of information flow? Returning to ”decoding the atmosphere” and reduce the problem to 1s and 0s, or symbols. If we treat the atmosphere as a perfect encoder, with the message being the phenomenon of interest, then we decode this progressively through the ”information cascade”. (Weiner or radar detection is similar, where only noise can be reduced and not the signal. Likewise, we have finite resources to decode the atmosphere. How do we optimise this? For instance, bias-variance tradeoff (aka discrimination versus reliability/calibration)

IG framework and interpreting results (bits v skill score; do the number of bits have a meaning?) What are the advantages of using IG framework?

a. Reliability diagrams

Create four sets of fake data: high/low discrimination (many obs close to climatology vs. many far away) and high/low reliability (25pc of 0.2 precip forecasts verify, versus 2pc). Do interpolated line and apply 1-5 system. Also find way to doing same for discrimination.

Label the uncertainty and XESS areas. What else? How to state ”error is due to X and Y” or in underfit/overfit terms?

We are forecasting a supercell, 10pc chance over numerous days.

Good REL, good DSC. If the forecaster gets as close to zero and unity as possible when issuing probabilities, and is correct sufficiently often, both relability and discrimination is high. Of course, information from discrimination can only be garnered if the forecasts deviate from the single climatology number.

Good REL, poor DSC. If the forecaster stays close to 10pc, it may be that a supercell was observed 10pc of the time. There, there will be good reliability on average, but the summation of error over each time shows an accumulation of error from such little correspondence between the likelihood and the event occurring more often than not (poor DSC) stemming from the lack of deviation from climate.

Poor REL, good DSC. If the forecaster is highly confident, they may go very close to yes/no certainty and be correct sufficiently often to be skilful on average. When we break down that score into REL and DSC, we find that information gained from the times the forecast was correct was larger than the information lost after highly certain forecasts. This is a high-risk strategy and is potentially disasterous for end-users with a low cost–loss ratio (i.e., it costs far less to mitigate an undesirable outcome than the loss incurred from that outcome). These users require the safety net of a larger reliability error (to error on the side of caution once a critical probabilitiy threshold has been reached) at the cost of avoiding the disaster of overconfidence. (also, asymmetrical preference of being surprised one way vs. other. Also, show the cost/loss surprise table 2x2?)

Poor REL, poor DSC. Let us say a spiteful forecaster were to sample their forecast probabilities from a narrow Gaussian distribution heavily skewed towards unity. This would

generate typically incorrect answers that tend to be worse than simply picking 0.5 probability, and skewed away from the climatological average. This is dangerous, not only useless, characterised by a cluster of random forecasts close to unity whilst the observed PDF is wider and close to 0.1.

b. Link with cost–loss ratio

Buizza showed forecast value.

In prob sense, Bayesian too? Show 2x2 table of surprise in bits per event. That can be decomposed after aggregation of cases, so we can total up each segment of the 2x2 table to see total incurred surprise (both good and bad!)

c. Application

Weisheimer and Palmer - creating ”1-5” scale of usefulness related to the three components. (Cost/loss ratio?) How to present.

How to connect with ”alpha” parameter or what controls the fitting of a ML model.

We need better obs uncertainties. The XES allows uncertainty to be changed on the fly with each calculation, while the model will not be punished as harshly as otherwise (without ob error) for a worse forecast during more uncertain observations (e.g., radar data that is occasionally attenuated).

The issue with rare events is that representative decomposition requires a sufficient sample size. Without this, it may be better to inspect the XES in bits case-by-case (?).

9. Conclusions

Blah

The benefit of using XES in rare events. As in Fig. XX in Benedetti (REF 2010), differences between using Brier-based scores and information-theoretical scores becomes starker as events become rarer. We show in intermittent regimes (such as in the Lorenz-63 system, which might represent tornado formation etc) that interpreting a model as skilful or not is sensitive to the score used. Even small differences can be important when tuning ML or deciding on a model upgrade, so how do we decide which is correct? The caveats of info theory aside, it is the logical better one: three reasons and ref (Benedetti).

Benefit of using error. The Brier Score likewise offers the intergration of model error. It is important because (show Weijs paper with ob error sensitivity).

Benefit of decomposition. See below about decomp.

Benefit of properties like unitless, proper, etc

The comparison between two forecast systems is only justifiable when the same time periods are analysed. Then, simply the XES can be used. The resulting bits show the amount of information that has been delivered to the user: if this is positive (as produced by Eqn XX), it show information gain; if negative, then information loss. This was demonstrated in (Lawson 2021 REF), where thunderstorms were predicted with a blanket prediction of 20%. (MORE HERE).

However, if two predictive systems are to be compared with each other, or if we want to see if a forecast system can outform simply picking the climatology (base rate or event frequency), we must use XESS. This normalises the information gained from the forecast by the uncertainty of the system in which it is forecasting. A positive XESS represents a better forecast than the denominator. The skill score can also be turned into skill scores likewise, creating DSCSS and RELSS. For instance, $DSCSS < RELSS > 0$ suggests the forecast is not

skillful because the spread (uncertainty) of the forecast system may not be calibrated to the uncertainty of the ovserved system. This is independent of UNC.

Other commonly used scores in meteorology are susceptible to the problem with (MEAN SQ ERROR?). The oft-used Fractions Skill Score (ref) and its probabilistic equivalent (Duc eFSS REF) use the (MSE?) Brier Score (in denom in eq. XX in which?). Swapping the (FBS?) for an information-theory analogue (such as DKL/IGN? Eqn.XX ?) means the interpretation of the information-theoretical analogue would be similar, but with the benefits outlined above.

We conceptualise this as an information dam from naïvity to best guess (climatology) to post-NWP guidance to forecast issuance to interpretation by decision makers (or the lay person). At each point, there is an adjustment to the remaining ignorance (independent of the UNC that cannot be removed, merely negated), and this score decomposition can assess whether it is due to good sharpness, good calibration, or both. This then fits in machine learning (e.g., alpha; bias–variance tradeoff), deciding how to fix initial conditions in NWP/DA, evaluating human interpretation of uncertainty and confidence; and even alerting to situations where information coming in will take longer to interpret effectively (entropy is higher - more choices - UNC measures this - these equations assume optional decision making).

Link to jupyter notebook repo and classes to compute scores.

Potential appendix for derivations?

a. Future work

The authors strongly assert their desire to make theoretical concepts relevant to the operational and model-development community.

On the theoretical side, The information-theory framework allows things like K-S entropy (uncertainty growth like Lorenz's chaotic saturation etc). Predictability is lost - a time horizon is reached (palmer) - just like information transfer asymptotes to zero (mutual information becomes almost entirely redundant, like saying the same thing five times). Block codes - is this the analogue of increasing efficency to the point of total trust, then a mistake becomes catastrophic. Determinstic, high-res, curse of dimensionality issue... is this the meteorology version?

Also, bits v trits

On the practical side, Demonstrating a "fractional informational gain" where the ensemble Fractions Skill Score (eFSS) is modified to use XES instead of (MAE? BS? MSE?)

Acknowledgments. The authors thank...

APPENDIX

Derivations

a. BS is a second-order approximation of XES/DKL

Can't make it work!

b. Information Gain

Can show various conversions between quantities, and how to decompose the equations?

REFERENCES

LIST OF TABLES

TABLE 1. Analogues of jargon between multiple fields

Calibration	Reliability, spread	Variance	Symbols
Redundancy, Mutual information	Discrimination, Resolution	Bias	DSC

TABLE 2. Lists of Experiments using L63 model.

Window size	1, 2, ??
IC error	1e-5, 1e-6, ??
Intermittency factor	11, 12, ??
Ensemble size	10, 100, ?