Verification of Numerical Weather Prediction (NWP) Forecasts

Roland Stull
Geophysical Disaster Computational Fluid Dynamics Center
Earth, Ocean & Atmos. Science Dept.
University of British Columbia
2020-2207 Main mall
Vancouver, BC V6T 1Z4 Canada

rstull @ eos.ubc.ca

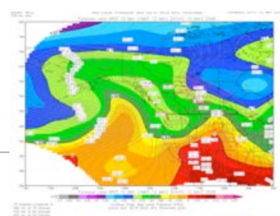
Special Thanks to my NWP Team:
Henryk Modzelewski, George Hicks II, Thomas Nipen, Dominique Bourdin, Atoossa Bakhshaii, May Wong, Bruce Thomson, Greg West
& many former team members.

Verification = a measure of Forecast Quality


wikipedia commons

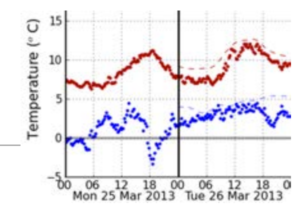
- Accuracy (a BAD measure)
- Skill (accuracy relative to some reference such as:
  - climatology (averaged over 30 years)
  - persistence (same as previous weather)
  - random (Monte-Carlo, bootstrapping, etc.)

Verification Topics



- Continuous Variables [T, U, V, P, bounded (RH, Precip)...]
- Categorical (Binary, yes/no) Events
- Probabilistic & Ensemble Forecasts
- Terrain Issues

Continuous Variables



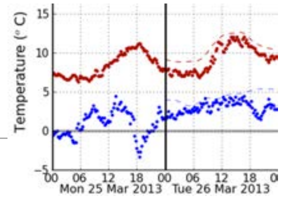
- Define variables

$$A = \text{initial analysis (based on observations)}$$
$$V = \text{verifying analysis (based on later obs.)}$$
$$F = \text{deterministic forecast}$$
$$C = \text{climatological conditions}$$
$$n = \text{number of grid points being averaged}$$

- Anomaly = Difference from climatology

$$F - C = \text{predicted anomaly}$$
$$A - C = \text{persistence anomaly}$$
$$V - C = \text{verifying anomaly}$$

Continuous Variables



- Tendency = change with time

$$F - A = \text{predicted tendency}$$
$$V - A = \text{verifying tendency}$$

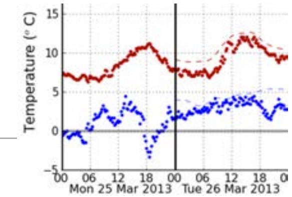
- Error = difference from observations or from verifying analysis

$$F - V = \text{forecast error}$$
$$A - V = \text{persistence error}$$

- Average (defined) for any variable X. (k = time or grid index) over n times, over n grid points, or both:

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k \qquad \bullet(20.24)$$

Continuous Variables



- Mean Error (ME) = bias = systematic error

$$ME = \overline{(F - V)} = \bar{F} - \bar{V} \qquad -\infty \text{ to } \infty.\ \ 0 = \text{best}$$

- Mean Absolute Error (MAE)

$$MAE = \overline{|F - V|} \qquad 0 \text{ to } \infty.\ \ 0 = \text{best}$$

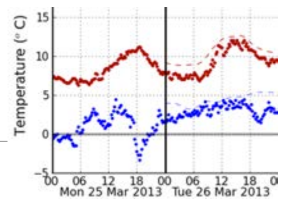
- Mean Squared Error (MSE) and Root Mean Sq. Error (RMSE)

$$MSE = \overline{(F - V)^2} \qquad 0 \text{ to } \infty.\ \ 0 = \text{best}$$

$$RMSE = \sqrt{\overline{(F - V)^2}} \qquad 0 \text{ to } \infty.\ \ 0 = \text{best}$$

Continuous Variable **Skill Scores**

$$SkillScore = \frac{score - ref.}{perfect - ref.}$$



- Mean Squared Error Skill Score (MSESS)

$$MSESS = 1 - \frac{MSE}{MSE_{ref}} \qquad -\infty \text{ to } 1.\ \ 1 = \text{best}$$

(0=no better than ref.)

- Root Mean Squared Error Skill Score (RMSESS)

$$RMSESS = 1 - \frac{RMSE}{RMSE_{ref}} \qquad -\infty \text{ to } 1.\ \ 1 = \text{best}$$

(0=no better than ref.)

- Mean Absolute Error Skill Score (MAESS)

$$MAESS = 1 - \frac{MAE}{MAE_{ref}} \qquad -\infty \text{ to } 1.\ \ 1 = \text{best}$$

(0=no better than ref.)

where persistence or climate is often used as the reference

Continuous **Bounded** Variable: Precipitation

- Bias Ratio (BR) = systematic error

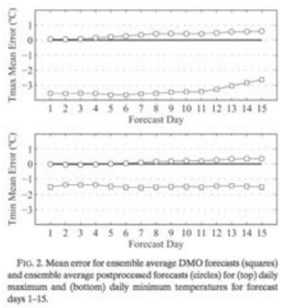
$$BR = \overline{\left(\frac{F}{V}\right)} \qquad 0 \text{ to } \infty.\ \ 1 = \text{best}$$

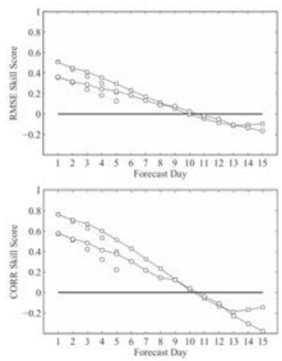
- Degree of Mass Balance (DMB)

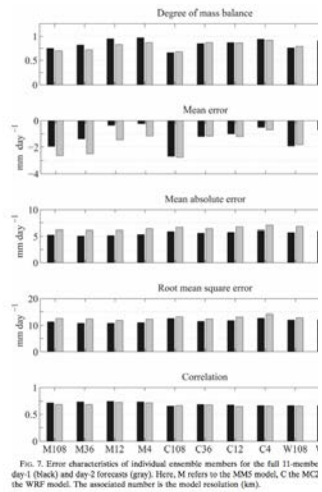
$$DMB = \frac{\bar{F}}{\bar{V}} \qquad 0 \text{ to } \infty.\ \ 1 = \text{best}$$

- ..

Examples

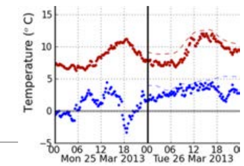






papers by McCollor & Stull

Continuous Variables



- Pearson product-moment **Correlation Coefficient** (r):

$$r = \frac{\overline{F'V'}}{\sqrt{\overline{(F')^2}} \cdot \sqrt{\overline{(V')^2}}}$$

–1 to 1. 1 = best
(–1 = F varies opposite to V)

$$F' = F - \overline{F} \quad \text{and} \quad V' = V - \overline{V}$$

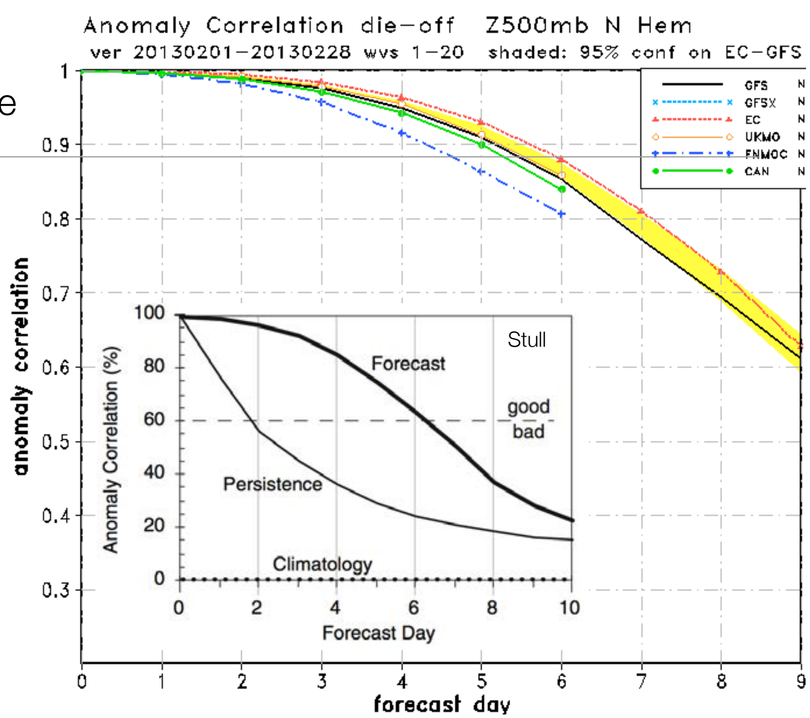
- Anomaly Correlation (AC):

–1 to 1. 1 = best
(0=no better than clim.)

$$\text{AC} = \frac{\overline{\left[(F-C)-\overline{(F-C)}\right]\cdot\left[(V-C)-\overline{(V-C)}\right]}}{\sqrt{\overline{\left[(F-C)-\overline{(F-C)}\right]^2} \cdot \overline{\left[(V-C)-\overline{(V-C)}\right]^2}}}$$

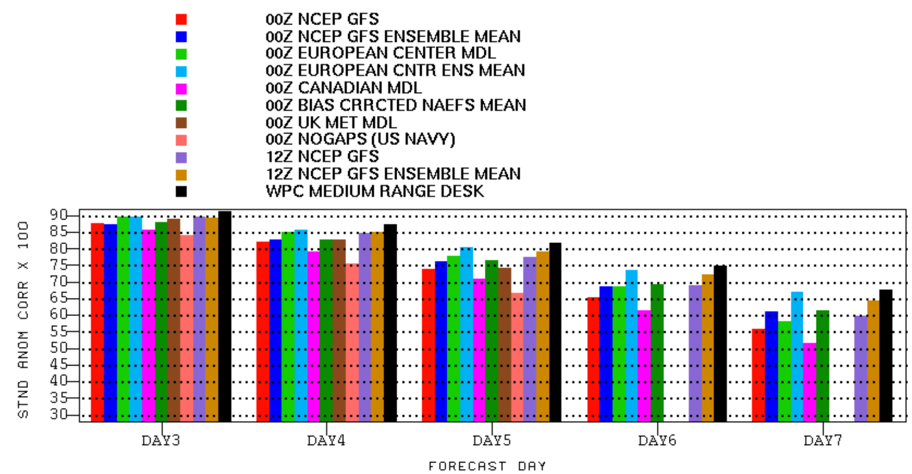
Example

NCEP



Example

CONUS REGION 12 UTC PMSL FRCST VERIFICATION FOR 20120326 THRU 20130326



Solved Example (§)

Given the following synthetic analysis (*A*), NWP forecast (*F*), verification (*V*), and climate (*C*) fields of 50-kPa height (km). Each field represents a weather map (North at top, East at right).

Analysis:

5.3	5.3	5.3	5.4
5.4	5.3	5.4	5.5
5.5	5.4	5.5	5.6
5.6	5.5	5.6	5.7
5.7	5.6	5.7	5.7

Forecast:

5.5	5.2	5.2	5.3
5.6	5.4	5.3	5.4
5.6	5.5	5.4	5.5
5.7	5.6	5.5	5.6
5.7	5.7	5.6	5.6

Verification:

5.4	5.3	5.3	5.3
5.5	5.4	5.3	5.4
5.5	5.5	5.4	5.5
5.6	5.6	5.5	5.6
5.6	5.7	5.6	5.7

Climate:

5.4	5.4	5.4	5.4
5.4	5.4	5.4	5.4
5.5	5.5	5.5	5.5
5.6	5.6	5.6	5.6
5.7	5.7	5.7	5.7

Find the mean error of the forecast and of persistence. Find the forecast MAE and MSE. Find MSEC and MSESS. Find the correlation coefficient between the forecast and verification. Find the RMS errors and the anomaly correlations for the forecast and persistence.

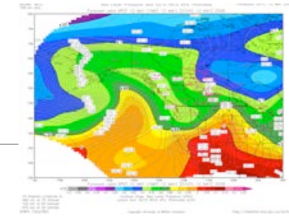
Solution

Use eq. (20.25): $ME_{forecast} = 0.01$ km $=$ **10 m**
Use eq. (20.26): $ME_{persistence} =$ **15 m**
Use eq. (20.27): $MAE =$ **40 m**
Use eq. (20.28): $MSE_{forecast} = 0.004$ km^2 $=$ **4000 m^2**
Use eq. (20.28): $MSEC =$ 0.0044 km^2 $=$ **4500 m^2**
Use eq. (20.29): $MSESS = 1 - (4000/4500) =$ **0.11**
Use eq. (20.30): $RMSE_{forecast} =$ **63 m**
Use eq. (20.30): $RMSE_{persistence} =$ **87 m**
Use eq. (20.31): $r =$ **0.92** (dimensionless)
Use eq. (20.33): forecast anomaly correlation = **81.3%**
Use eq. (20.34): persist. anomaly correlation = **7.7%**

Check: Units OK. Physics OK.

Discussion: Analyze (i.e., draw height contour maps for) the analysis, forecast, verification, and climate fields. The analysis shows a Rossby wave with ridge and trough, and the verification shows this wave moving east. The forecast amplifies the wave too much. The climate field just shows the average of higher heights to the south and lower heights to the north, with all transient Rossby waves averaged out.

Verification Topics



- Continuous Variables [T, U, V, P, bounded (RH, Precip)...]
- Categorical (Binary, yes/no) Events
- Probabilistic & Ensemble Forecasts
- Terrain Issues

Categorical (yes/no)

- True binary:
  - Snow / no-snow
  - Rain / no-rain
  - Sun / shade
- Threshold exceedence
  - $T < T_{threshold}$
  - Precip > Precip.$_{threshold}$
  - Wind > Wind.$_{threshold}$
  - etc.

Contingency Table

(a)

		Observation: Yes	Observation: No
Forecast	Yes	Hit	False Alarm
Forecast	No	Miss	Correct Rejection or Correct Negative

(b)

		Observation: Yes	Observation: No
Forecast	Yes	a	b
Forecast	No	c	d

(c)

		Observation: Yes	Observation: No
Forecast	Yes	Mitigated Loss (C)	Cost (C)
Forecast	No	Loss (L)	No Cost (0)

Figure 20.23
Contingency table for a binary (Yes/No) situation. "Yes" means the event occurred or was forecast to occur. (a) Meaning of cells. (b) Counts of occurrences, where a + b + c + d = n. (c) Expense matrix, where C = cost for taking protective action (i.e., for mitigating the loss), and L = loss due to an unmitigated event.

Binary Scores

The **bias score** *B* indicates over- or under-prediction of the frequency of event occurrence:

$$B = \frac{a+b}{a+c} \qquad (20.36)$$

The **portion correct** *PC* (also known as **portion of forecasts correct** *PFC*) is

$$PC = \frac{a+d}{n} \qquad (20.37)$$

But perhaps a portion of PC could have been due to random-chance (dumb but lucky) forecasts. Let *E* be this "random luck" portion, assuming that you made the same ratio of "YES" to "NO" forecasts:

$$E = \left(\frac{a+b}{n}\right)\cdot\left(\frac{a+c}{n}\right) + \left(\frac{d+b}{n}\right)\cdot\left(\frac{d+c}{n}\right) \qquad (20.38)$$

We can now define the portion of correct forecasts that was actually skillful (i.e., not random chance), which is known as the **Heidke skill score** (*HSS*):

$$HSS = \frac{PC - E}{1 - E} \qquad (20.39)$$

Binary Scores

The **hit rate** *H* is the portion of actual occurrences (obs. = "YES") that were successfully forecast:

$$H = \frac{a}{a+c} \qquad \bullet(20.40)$$

It is also known as the **probability of detection** *POD*.

The **false-alarm rate** *F* is the portion of non-occurrences (observation = "NO") that were incorrectly forecast:

$$F = \frac{b}{b+d} \qquad \bullet(20.41)$$

Don't confuse this with the **false-alarm ratio** *FAR*, which is the portion of "YES" forecasts that were wrong:

$$FAR = \frac{b}{a+b} \qquad (20.42)$$

Binary Scores

A **true skill score** *TSS* (also known as **Peirce's skill score** *PSS*, and as **Hansen and Kuipers' score**) can be defined as

$$TSS = H - F \qquad (20.43)$$

which is a measure of how well you can discriminate between an event and a non-event, or a measure of how well you can detect an event.

A **critical success index** *CSI* (also known as a **threat score** *TS*) is:

$$CSI = \frac{a}{a+b+c} \qquad (20.44)$$

Binary Scores

Suppose we consider the portion of hits that might have occurred by random chance a_r:

$$a_r = \frac{(a+b)\cdot(a+c)}{n} \qquad (20.45)$$

Then we can subtract this from the actual hit count to modify CSS into an **equitable threat score** *ETS*, also known as **Gilbert's skill score** *GSS*:

$$GSS = \frac{a - a_r}{a - a_r + b + c} \qquad (20.46)$$

which is also useful for rare events.

For a perfect forecasts (where $b = c = 0$), the values of these scores are: $B = 1$, $PC = 1$, $HSS = 1$, $H = 1$, $F = 0$, $FAR = 0$, $TSS = 1$, $CSS = 1$, $GSS = 1$.

For totally wrong forecasts (where $a = d = 0$): $B = 0$ to ∞, $PC = 0$, $HSS =$ negative, $H = 0$, $F = 1$, $FAR = 1$, $TSS = -1$, $CSS = 0$, $GSS =$ negative.

Binary Scores

Solved Example

Given the following contingency table, calculate all the binary verification statistics.

		Observation	
		Yes	No
Forecast	Yes:	90	50
	No	75	150

Solution:

Given: $a = 90$, $b = 50$, $c = 75$, $d = 150$

Find: *B, PC, HSS, H, F, FAR, TSS, CSI, GSS*

First, use eq. (20.35): $n = 90 + 50 + 75 + 150 = 365$

So apparently we have daily observations for a year.

Use eq. (20.36): $B = (90 + 50) / (90 + 75) = \underline{\mathbf{0.85}}$

Use eq. (20.37): $PC = (90 + 150) / 365 = \underline{\mathbf{0.66}}$

Use eq. (20.38):

$E = [(90{+}50)\cdot(90{+}75) + (150{+}50)\cdot(150{+}75)]/(365^2)$

$E = 68100 / 133225 = \underline{\mathbf{0.51}}$

Use eq. (20.39): $HSS = (0.66 - 0.51) / (1 - 0.51) = \underline{\mathbf{0.31}}$

Use eq. (20.40): $H = 90 / (90 + 75) = \underline{\mathbf{0.55}}$

Use eq. (20.41): $F = 50 / (50 + 150) = \underline{\mathbf{0.25}}$

Use eq. (20.42): $FAR = 50 / (90 + 50) = \underline{\mathbf{0.36}}$

Use eq. (20.43): $TSS = 0.55 - 0.25 = \underline{\mathbf{0.30}}$

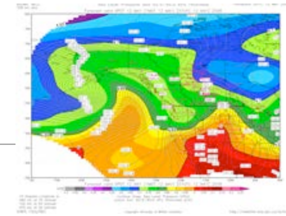
Use eq. (20.44): $CSI = 90 / (90 + 50 + 75) = \underline{\mathbf{0.42}}$

Use eq. (20.45): $a_r = [(90{+}50)\cdot(90{+}75)]/365 = 63.3$

Use eq. (20.46):

$GSS = (90{-}63.3) / (90{-}63.3{+}50{+}75) = \underline{\mathbf{0.18}}$

Verification Topics



- Continuous Variables [T, U, V, P, bounded (RH, Precip)...]
- Categorical (Binary, yes/no) Events
- Probabilistic & Ensemble Forecasts
- Terrain Issues

Probability Fcst Verification

Brier Skill Score

For calibrated probability forecasts, a **Brier skill score** (*BSS*) can be defined relative to climatology as

$$BSS = 1 - \frac{\sum_{k=1}^{N} (p_k - o_k)^2}{\left(\sum_{k=1}^{N} o_k\right)\cdot\left(N - \sum_{k=1}^{N} o_k\right)} \qquad \bullet(20.47)$$

where p_k is the forecast probability ($0 \le p_k \le 1$) that the threshold will be exceeded (e.g., the probability that the precipitation will exceed a precipitation threshold) for any one forecast *k*, and *N* is the number of forecasts. The verifying observation $o_k = 1$ if the observation exceeded the threshold, and is set to zero otherwise.

BSS = 0 for a forecast no better than climatology. *BSS* = 1 for a perfect deterministic forecast (i.e., the forecast is $p_k = 1$ every time the event happens, and $p_k = 0$ every time it does not). For probabilistic forecasts, $0 \le BSS \le 1$. Larger BSS values are better.

Probability Fcst Verification

Reliability

How **reliable** are the probability forecasts? Namely, when we forecast an event with a certain probability, is it observed with the same relative frequency? To determine this, after you make each forecast, sort it into a forecast probability bin (*j*) of probability width Δp, and keep a tally of the number of forecasts (n_j) that fell in this bin, and count how many of the forecasts verified (n_{oj}, for which the corresponding observation satisfied the threshold).

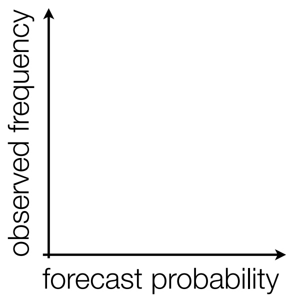
For example, if you use bins of size $\Delta p = 0.1$, then create a table such as:

bin index	bin center	fcst. prob. range	n_j	n_{oj}
$j = 0$	$p_j = 0$	$0 \le p_k < 0.05$	n_0	n_{o0}
$j = 1$	$p_j = 0.1$	$0.05 \le p_k < 0.15$	n_1	n_{o1}
$j = 2$	$p_j = 0.2$	$0.15 \le p_k < 0.25$	n_2	n_{o2}
etc.	...	...	...	...
$j = 9$	$p_j = 0.9$	$0.85 \le p_k < 0.95$	n_9	n_{o9}
$j = 10 = J$	$p_j = 1.0$	$0.95 \le p_k \le 1.0$	n_{10}	n_{o10}

A plot of the observed relative frequency (n_{oj}/n_j) on the ordinate vs. the corresponding forecast probability bin center (p_j) on the abscissa is called a **reliability diagram**. For perfect reliability, all the points should be on the 45° diagonal line.

Ensemble Fcst Verification

Reliability Diagram example



MCCOLLOR AND STULL

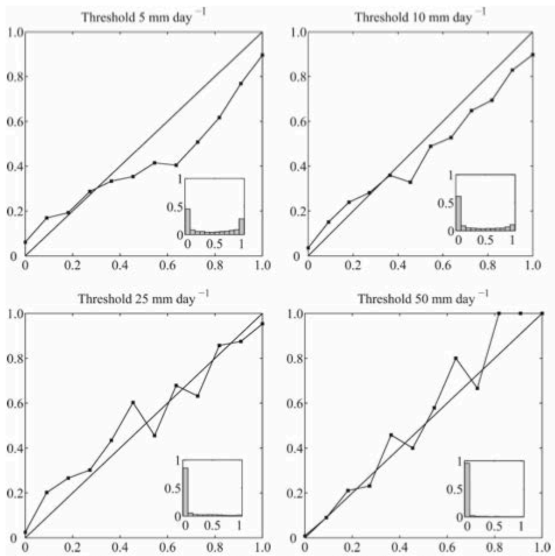


FIG. 12. Reliability diagrams from the full 11-member ensemble for the day-1 forecasts for precipitation thresholds of 5, 10, 25, and 50 mm day^{-1}. The abscissas in these graphs show the forecast probability, and the ordinates are the observed relative frequency. The straight diagonal line indicates perfect reliability. The inset diagrams depict sharpness, where the abscissa is forecast probability and the ordinate is the relative frequency of use.

Probability Fcst Verification

Reliability Diagram

Ideally, the probabilistic forecast–observation points lie on the diagonal of the reliability diagram, indicating the event is always forecast at the same frequency it is observed. The reliability component of the Brier score in a graphical representation is the weighted, averaged, squared distance between the reliability curve and the 45° diagonal line. If the points lie above (below) the diagonal, it means the event is underforecast (overforecast). Reliability curves with a zigzag shape centered on the diagonal indicate good reliability represented by a small sample size. Poor reliability can be improved substantially by appropriate a posteriori calibration and/or postprocessing of forecasts delivered from an established system, though it is a difficult task to achieve in a real-time operational EPS (Atger 2003).

Probability Fcst Verification

A Brier skill score for **relative reliability** ($BSS_{reliability}$) is:

$$BSS_{reliability} = \frac{\sum_{j=0}^{J}\left[(n_j \cdot p_j) - n_{oj}\right]^2}{\left(\sum_{k=1}^{N} o_k\right)\cdot\left(N - \sum_{k=1}^{N} o_k\right)} \qquad (20.48)$$

where $BSS_{reliability} = 0$ for a perfect forecast.

Probability Fcst Verification

Solved Example (§)

Given the table below of $k = 1$ to 31 forecasts of the probability p_k that the temperature will be below threshold 20°C, and the verification $o_k = 1$ if indeed the observed temperature was below the threshold.

(a) Find the Brier skill score. (b) For probability bins of width $\Delta p = 0.2$, plot a reliability diagram, and find the reliability Brier skill score.

k	p_k	o_k	BN	j	k	p_k	o_k	BN	j
1	0.43	0	0.18	2	16	0.89	1	0.01	4
2	0.98	1	0.00	5	17	0.13	0	0.02	1
3	0.53	1	0.22	3	18	0.92	1	0.01	5
4	0.33	1	0.45	2	19	0.86	1	0.02	4
5	0.50	0	0.25	3	20	0.90	1	0.01	5
6	0.03	0	0.00	0	21	0.83	0	0.69	4
7	0.79	1	0.04	4	22	0.00	0	0.00	0
8	0.23	0	0.05	1	23	1.00	1	0.00	5
9	0.20	1	0.64	1	24	0.69	0	0.48	3
10	0.59	1	0.17	3	25	0.36	0	0.13	2
11	0.26	0	0.07	1	26	0.56	1	0.19	3
12	0.76	1	0.06	4	27	0.46	0	0.21	2
13	0.17	0	0.03	1	28	0.63	0	0.40	3
14	0.30	0	0.09	2	29	0.10	0	0.01	1
15	0.96	1	0.00	5	30	0.40	1	0.36	2
					31	0.73	1	0.07	4

Solution

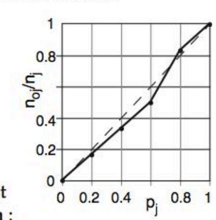
Given: The white portion of the table above.
Find: BSS = ? , $BSS_{reliability}$ = ?, and plot reliability.

(a) Use eq. (20.47). The grey-shaded column labeled BN shows each contribution to the numerator $(p_k - o_k)^2$ in that eq. The sum of BN = 4.86. The sum of o_k = 16. Thus, the eq is: $BSS = 1 - [4.86 / \{16 \cdot (31–16)\}] =$ **0.98**

(b) There are $J = 6$ bins, with bin centers at p_j = 0, 0.2, 0.4, 0.6, 0.8, and 1.0. (Note, the first and last bins are one-sided, half-width relative to the nominal "center" value.) I sorted the forecasts into bins using j = round($p_k/\Delta p$, 0), giving the grey j columns above.

For each j bin, I counted the number of forecasts n_j falling in that bin, and I counted the portion of those forecasts that verified n_{oj}. See table below:

j	p_j	n_j	n_{oj}	n_{oj}/n_j	num
0	0	2	0	0	0
1	0.2	6	1	0.17	0.04
2	0.4	6	2	0.33	0.16
3	0.6	6	3	0.50	0.36
4	0.8	6	5	0.83	0.04
5	1.0	5	5	1	0



The observed relative frequency n_{oj}/n_j plotted against p_j is the **reliability diagram** :

Use eq. (20.48). The contribution to the numerator from each bin is in the num column above, which sums to 0.6 . Thus: $BSS_{reliability} = 0.6/ \{16 \cdot (31–16)\} =$ **0.0025**

Probability Fcst Verification

b. Continuous ranked probability score

The continuous ranked probability score (CRPS; Hersbach 2000) is a complete generalization of the Brier score. The Brier score (Brier 1950; Atger 2003; McCollor

The CRPS can be interpreted as an integral over all possible Brier scores. A major advantage of the Brier score is that it can be decomposed into a reliability component, a resolution component, and an uncertainty component (Murphy 1973; Toth et al. 2003). In a similar fashion, Hersbach (2000) showed how the CRPS can be decomposed into the same three components:

$$\mathrm{CRPS} = \mathrm{Reli} - \mathrm{Resol} + \mathrm{Unc}. \qquad (2)$$

Probability Fcst Verification

Continuous Ranked Probability Skill Score

The CRPS components can be converted to a positively oriented skill score (CRPSS) in the same manner that the Brier score components are converted to a skill score (BSS; McCollor and Stull 2008b):

$$\begin{aligned}\mathrm{CRPSS} &= \frac{\mathrm{Resol}}{\mathrm{Unc}} - \frac{\mathrm{Reli}}{\mathrm{Unc}} \\ &= \text{relative resolution} - \text{relative reliability} \\ &= \mathrm{CRPS}_{\mathrm{RelResol}} - \mathrm{CRPS}_{\mathrm{RelReli}}.\end{aligned}$$

Probability Fcst Verification

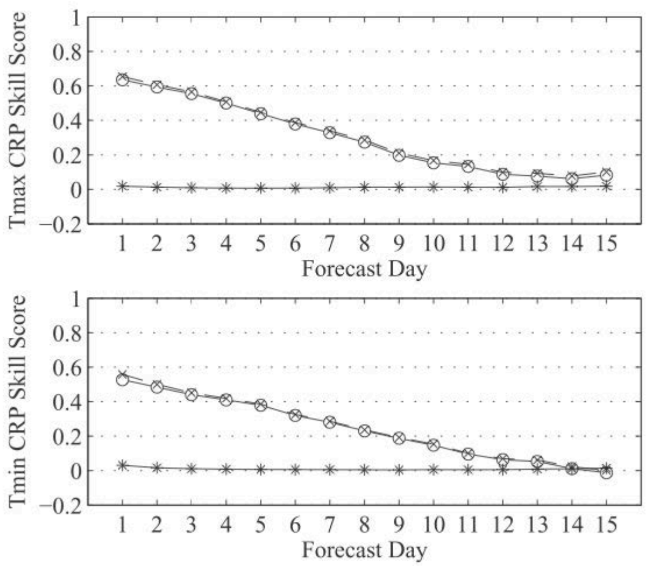


FIG. 4. Continuous ranked probability skill scores for ensemble (top) daily maximum and (bottom) daily minimum temperature anomaly forecasts. The CRP skill score (circles) equals the relative resolution (x markers) less the relative reliability (asterisks).

Probability Fcst Verification

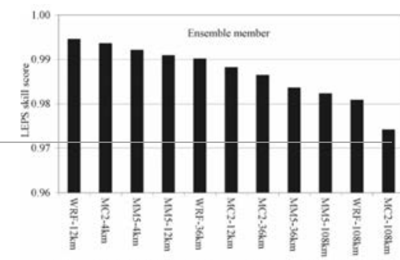


FIG. 6. The LEPS skill score for ensemble members for the day-2 forecasts. Values closer to 1 are better. The coarsest-resolution members show a clear deterioration of skill compared to the finer-resolution members, as for the day-1 forecasts. In general, the day-2 forecasts are slightly less skillful than are the day-1 forecasts in terms of the LEPS skill score.

- Linear Error in Probability Space (LEPS)

Linear error in probability space (LEPS) is defined as the mean absolute difference between the cumulative frequency of the forecasts and the cumulative frequency of the observations (Déqué 2003). LEPS ensures that error in the center of the distribution is treated with more importance than error found in the extreme tail of the distribution. A LEPS skill score can be defined with the climatological median as a reference (Wilks 1995).

ROC diagram

ROC Diagram

A **Relative Operating Characteristic** (ROC) diagram shows how well a probabilistic forecast can **discriminate** between an event and a non-event. For example, an event could be heavy rain that causes flooding, or cold temperatures that cause crops to freeze. The probabilistic forecast could come from an ensemble forecast, as illustrated next.

ROC diagram

Suppose that the individual NWP models of an $N = 10$ member ensemble made the following forecasts of 24-h accumulated rainfall R for Day 1:

NWP model	R (mm)	NWP model	R (mm)
Model 1	8	Model 6	4
Model 2	10	Model 7	20
Model 3	6	Model 8	9
Model 4	12	Model 9	5
Model 5	11	Model 10	7

Consider a precipitation threshold of 10 mm. The ensemble above has 4 models that forecast 10 mm or more, hence the forecast probability is $p_1 = 4/N = 4/10 = 40\%$. Supposed that 10 mm or more of precipitation was indeed observed, so the observation flag is set to one: $o_1 = 1$.

ROC diagram

On Day 2, three of the 10 models forecast 10 mm or more of precipitation, hence the forecast probability is $p_2 = 3/10 = 30\%$. On this day precipitation did NOT exceed 10 mm, so the observation flag is set to zero: $o_2 = 0$. Similarly, for Day 3 suppose the forecast probability is $p_3 = 10\%$, but heavy rain was observed, so $o_3 = 1$. After making ensemble forecasts every day for a month, suppose the results are as listed in the left three columns of Table 20-4.

Table 20-4. Sample calculations for a ROC diagram. Forecast flags f are shown under the probability thresholds.

Day	o	p (%)	Probability Threshold $p_{threshold}$ (%): 0	10	20	30	40	50	60	70	80	90	100
1	1	40	1	1	1	1	1	0	0	0	0	0	0
2	0	30	1	1	1	1	0	0	0	0	0	0	0
3	1	10	1	1	0	0	0	0	0	0	0	0	0
4	1	50	1	1	1	1	1	1	0	0	0	0	0
5	0	60	1	1	1	1	1	1	1	0	0	0	0
6	0	30	1	1	1	1	0	0	0	0	0	0	0
7	0	40	1	1	1	1	1	0	0	0	0	0	0
8	1	80	1	1	1	1	1	1	1	1	1	0	0
9	0	50	1	1	1	1	1	1	0	0	0	0	0
10	1	20	1	1	1	0	0	0	0	0	0	0	0
11	1	90	1	1	1	1	1	1	1	1	1	1	0
12	0	20	1	1	1	0	0	0	0	0	0	0	0
13	0	10	1	1	0	0	0	0	0	0	0	0	0
14	0	10	1	1	0	0	0	0	0	0	0	0	0
15	1	70	1	1	1	1	1	1	1	1	0	0	0
16	0	70	1	1	1	1	1	1	1	1	0	0	0
17	1	60	1	1	1	1	1	1	1	0	0	0	0
18	1	90	1	1	1	1	1	1	1	1	1	1	0
19	1	80	1	1	1	1	1	1	1	1	1	0	0
20	0	80	1	1	1	1	1	1	1	1	1	0	0
21	0	20	1	1	1	0	0	0	0	0	0	0	0
22	0	10	1	1	0	0	0	0	0	0	0	0	0
23	0	0	1	0	0	0	0	0	0	0	0	0	0
24	0	0	1	0	0	0	0	0	0	0	0	0	0
25	1	70	1	1	1	1	1	1	1	1	0	0	0
26	0	10	1	1	0	0	0	0	0	0	0	0	0
27	0	0	1	0	0	0	0	0	0	0	0	0	0
28	1	90	1	1	1	1	1	1	1	1	1	1	0
29	0	20	1	1	1	0	0	0	0	0	0	0	0
30	1	80	1	1	1	1	1	1	1	1	1	0	0
Contingency Table Values		$a =$	13	13	12	11	11	10	9	8	6	3	0
		$b =$	17	14	10	7	5	4	3	2	1	0	0
		$c =$	0	0	1	2	2	3	4	5	7	10	13
		$d =$	0	3	7	10	12	13	14	15	16	17	17

Probability Fcst Verification

ROC diagram

An end user might need to make a decision to take action. Based on various economic or political reasons, the user decides to use a probability threshold of $p_{threshold}$ = 40%; namely, if the ensemble model forecasts a 40% or greater chance of daily rain exceeding 10 mm, then the user will take action. So we can set forecast flag f = 1 for each day that the ensemble predicted 40% or more probability, and f = 0 for the other days. These forecast flags are shown in Table 20-4 under the $p_{threshold}$ = 40% column.

Table 20-4. Sample calculations for a ROC diagram. Forecast flags f are shown under the probability thresholds.

Day	o	p (%)	Probability Threshold $p_{threshold}$ (%)										
			0	10	20	30	40	50	60	70	80	90	100
1	1	40	1	1	1	1	1	0	0	0	0	0	0
2	0	30	1	1	1	1	0	0	0	0	0	0	0
3	1	10	1	1	0	0	0	0	0	0	0	0	0
4	1	50	1	1	1	1	1	1	0	0	0	0	0
5	0	60	1	1	1	1	1	1	1	0	0	0	0
6	0	30	1	1	1	1	0	0	0	0	0	0	0
7	0	40	1	1	1	1	1	0	0	0	0	0	0
8	1	80	1	1	1	1	1	1	1	1	1	0	0
9	0	50	1	1	1	1	1	1	0	0	0	0	0
10	1	20	1	1	1	0	0	0	0	0	0	0	0
11	1	90	1	1	1	1	1	1	1	1	1	1	0
12	0	20	1	1	1	0	0	0	0	0	0	0	0
13	0	10	1	1	0	0	0	0	0	0	0	0	0
14	0	10	1	1	0	0	0	0	0	0	0	0	0
15	1	70	1	1	1	1	1	1	1	1	0	0	0
16	0	70	1	1	1	1	1	1	1	1	0	0	0
17	1	60	1	1	1	1	1	1	1	0	0	0	0
18	1	90	1	1	1	1	1	1	1	1	1	1	0
19	1	80	1	1	1	1	1	1	1	1	1	0	0
20	0	80	1	1	1	1	1	1	1	1	1	0	0
21	0	20	1	1	1	0	0	0	0	0	0	0	0
22	0	10	1	1	0	0	0	0	0	0	0	0	0
23	0	0	1	0	0	0	0	0	0	0	0	0	0
24	0	0	1	0	0	0	0	0	0	0	0	0	0
25	1	70	1	1	1	1	1	1	1	1	0	0	0
26	0	10	1	1	0	0	0	0	0	0	0	0	0
27	0	0	1	0	0	0	0	0	0	0	0	0	0
28	1	90	1	1	1	1	1	1	1	1	1	1	0
29	0	20	1	1	1	0	0	0	0	0	0	0	0
30	1	80	1	1	1	1	1	1	1	1	1	0	0
Contingency Table Values		a =	13	13	12	11	11	10	9	8	6	3	0
		b =	17	14	10	7	5	4	3	2	1	0	0
		c =	0	0	1	2	2	3	4	5	7	10	13
		d =	0	3	7	10	12	13	14	15	16	17	17

Probability Fcst Verification

ROC diagram

Other users might have other decision thresholds, so we can find the forecast flags for all the other probability thresholds, as given in Table 20-4. For an N member ensemble, there are only $(100/N) + 1$ discrete probabilities that are possible. For our example with N = 10 members, we can consider only 11 different probability thresholds: 0% (when no members exceed the rain threshold), 10% (when 1 out of the 10 members exceeds the threshold), 20% (etc.), . . . 90%, 100%.

Table 20-4. Sample calculations for a ROC diagram. Forecast flags f are shown under the probability thresholds.

Day	o	p (%)	Probability Threshold $p_{threshold}$ (%)										
			0	10	20	30	40	50	60	70	80	90	100
1	1	40	1	1	1	1	1	0	0	0	0	0	0
2	0	30	1	1	1	1	0	0	0	0	0	0	0
3	1	10	1	1	0	0	0	0	0	0	0	0	0
4	1	50	1	1	1	1	1	1	0	0	0	0	0
5	0	60	1	1	1	1	1	1	1	0	0	0	0
6	0	30	1	1	1	1	0	0	0	0	0	0	0
7	0	40	1	1	1	1	1	0	0	0	0	0	0
8	1	80	1	1	1	1	1	1	1	1	1	0	0
9	0	50	1	1	1	1	1	1	0	0	0	0	0
10	1	20	1	1	1	0	0	0	0	0	0	0	0
11	1	90	1	1	1	1	1	1	1	1	1	1	0
12	0	20	1	1	1	0	0	0	0	0	0	0	0
13	0	10	1	1	0	0	0	0	0	0	0	0	0
14	0	10	1	1	0	0	0	0	0	0	0	0	0
15	1	70	1	1	1	1	1	1	1	1	0	0	0
16	0	70	1	1	1	1	1	1	1	1	0	0	0
17	1	60	1	1	1	1	1	1	1	0	0	0	0
18	1	90	1	1	1	1	1	1	1	1	1	1	0
19	1	80	1	1	1	1	1	1	1	1	1	0	0
20	0	80	1	1	1	1	1	1	1	1	1	0	0
21	0	20	1	1	1	0	0	0	0	0	0	0	0
22	0	10	1	1	0	0	0	0	0	0	0	0	0
23	0	0	1	0	0	0	0	0	0	0	0	0	0
24	0	0	1	0	0	0	0	0	0	0	0	0	0
25	1	70	1	1	1	1	1	1	1	1	0	0	0
26	0	10	1	1	0	0	0	0	0	0	0	0	0
27	0	0	1	0	0	0	0	0	0	0	0	0	0
28	1	90	1	1	1	1	1	1	1	1	1	1	0
29	0	20	1	1	1	0	0	0	0	0	0	0	0
30	1	80	1	1	1	1	1	1	1	1	1	0	0
Contingency Table Values		a =	13	13	12	11	11	10	9	8	6	3	0
		b =	17	14	10	7	5	4	3	2	1	0	0
		c =	0	0	1	2	2	3	4	5	7	10	13
		d =	0	3	7	10	12	13	14	15	16	17	17

Probability Fcst Verification

ROC diagram

For each probability threshold, create a 2x2 contingency table with the elements a, b, c, and d as shown in Fig. 20.23b. For example, for any pair of observation and forecast flags (o_j, f_j) for Day j, use

a = count of days with hits $(o_j, f_j) = (1, 1)$.
b = count of days with false alarms $(o_j, f_j) = (0, 1)$.
c = count of days with misses $(o_j, f_j) = (1, 0)$.
d = count of days: correct rejection $(o_j, f_j) = (0, 0)$.

For our illustrative case, these contingency-table elements are shown near the bottom of Table 20-4.

Table 20-4. Sample calculations for a ROC diagram. Forecast flags f are shown under the probability thresholds.

Day	o	p (%)	Probability Threshold $p_{threshold}$ (%)										
			0	10	20	30	40	50	60	70	80	90	100
1	1	40	1	1	1	1	1	0	0	[illegible]	[illegible]	[illegible]	[illegible]
2	0	30	1	1	1	1	0	0	0	[illegible]	[illegible]	[illegible]	[illegible]
3	1	10	1	1	0	0	0	0	0	[illegible]	[illegible]	[illegible]	[illegible]
4	1	50	1	1	1	1	1	1	0	[illegible]	[illegible]	[illegible]	[illegible]
5	0	60	1	1	1	1	1	1	1	[illegible]	[illegible]	[illegible]	[illegible]
6	0	30	1	1	1	1	0	0	0	[illegible]	[illegible]	[illegible]	[illegible]
7	0	40	1	1	1	1	1	0	0	[illegible]	[illegible]	[illegible]	[illegible]
8	1	80	1	1	1	1	1	1	1	1	1	0	0
9	0	50	1	1	1	1	1	1	0	0	0	0	0
10	1	20	1	1	1	0	0	0	0	0	0	0	0
11	1	90	1	1	1	1	1	1	1	1	1	1	0
12	0	20	1	1	1	0	0	0	0	0	0	0	0
13	0	10	1	1	0	0	0	0	0	0	0	0	0
14	0	10	1	1	0	0	0	0	0	0	0	0	0
15	1	70	1	1	1	1	1	1	1	1	0	0	0
16	0	70	1	1	1	1	1	1	1	1	0	0	0
17	1	60	1	1	1	1	1	1	1	0	0	0	0
18	1	90	1	1	1	1	1	1	1	1	1	1	0
19	1	80	1	1	1	1	1	1	1	1	1	0	0
20	0	80	1	1	1	1	1	1	1	1	1	0	0
21	0	20	1	1	1	0	0	0	0	0	0	0	0
22	0	10	1	1	0	0	0	0	0	0	0	0	0
23	0	0	1	0	0	0	0	0	0	0	0	0	0
24	0	0	1	0	0	0	0	0	0	0	0	0	0
25	1	70	1	1	1	1	1	1	1	1	0	0	0
26	0	10	1	1	0	0	0	0	0	0	0	0	0
27	0	0	1	0	0	0	0	0	0	0	0	0	0
28	1	90	1	1	1	1	1	1	1	1	1	1	0
29	0	20	1	1	1	0	0	0	0	0	0	0	0
30	1	80	1	1	1	1	1	1	1	1	1	0	0
Contingency Table Values		a =	13	13	12	11	11	10	9	8	6	3	0
		b =	17	14	10	7	5	4	3	2	1	0	0
		c =	0	0	1	2	2	3	4	5	7	10	13
		d =	0	3	7	10	12	13	14	15	16	17	17

(b)

		Observation	
		Yes	No
Forecast	Yes	a	b
	No	c	d

Probability Fcst Verification

ROC diagram

Next, for each probability threshold, calculate the hit rate $H = a/(a+c)$ and false alarm rate $F = b/(b+d)$, as defined earlier in this chapter. These are shown in the last two rows of Table 20-4 for our example. When each (F, H) pair is plotted as a point on a graph, the result is called a **ROC diagram** (Fig. 20.24).

Table 20-4. Sample calculations for a ROC diagram. Forecast flags f are shown under the probability thresholds.

Day	o	p (%)	Probability Threshold $p_{threshold}$ (%)										
			0	10	20	30	40	50	60	70	80	90	100
1	1	40	1	1	1	1	1	0	0	0	0	0	0
2	0	30	1	1	1	1	0	0	0	0	0	0	0
3	1	10	1	1	0	0	0	0	0	0	0	0	0
4	1	50	1	1	1	1	1	1	0	0	0	0	0
5	0	60	1	1	1	1	1	1	1	0	0	0	0
6	0	30	1	1	1	1	0	0	0	0	0	0	0
7	0	40	1	1	1	1	1	0	0	0	0	0	0
8	1	80	1	1	1	1	1	1	1	1	1	0	0
9	0	50	1	1	1	1	1	1	0	0	0	0	0
10	1	20	1	1	1	0	0	0	0	0	0	0	0
11	1	90	1	1	1	1	1	1	1	1	1	1	0
12	0	20	1	1	1	0	0	0	0	0	0	0	0
13	0	10	1	1	0	0	0	0	0	0	0	0	0
14	0	10	1	1	0	0	0	0	0	0	0	0	0
15	1	70	1	1	1	1	1	1	1	1	0	0	0
16	0	70	1	1	1	1	1	1	1	1	0	0	0
17	1	60	1	1	1	1	1	1	1	0	0	0	0
18	1	90	1	1	1	1	1	1	1	1	1	1	0
19	1	80	1	1	1	1	1	1	1	1	1	0	0
20	0	80	1	1	1	1	1	1	1	1	1	0	0
21	0	20	1	1	1	0	0	0	0	0	0	0	0
22	0	10	1	1	0	0	0	0	0	0	0	0	0
23	0	0	1	0	0	0	0	0	0	0	0	0	0
24	0	0	1	0	0	0	0	0	0	0	0	0	0
25	1	70	1	1	1	1	1	1	1	1	0	0	0
26	0	10	1	1	0	0	0	0	0	0	0	0	0
27	0	0	1	0	0	0	0	0	0	0	0	0	0
28	1	90	1	1	1	1	1	1	1	1	1	1	0
29	0	20	1	1	1	0	0	0	0	0	0	0	0
30	1	80	1	1	1	1	1	1	1	1	1	0	0
Contingency Table Values		a =	13	13	12	11	11	10	9	8	6	3	0
		b =	17	14	10	7	5	4	3	2	1	0	0
		c =	0	0	1	2	2	3	4	5	7	10	13
		d =	0	3	7	10	12	13	14	15	16	17	17
Hit Rate: $H = a/(a+c)$ =			1.00	1.00	0.92	0.85	0.85	0.77	0.69	0.62	0.46	0.23	0.00
False Alarm Rate: $F=b/(b+d)$			1.00	0.82	0.59	0.41	0.29	0.26	0.18	0.12	0.06	0.00	0.00

ROC diagram

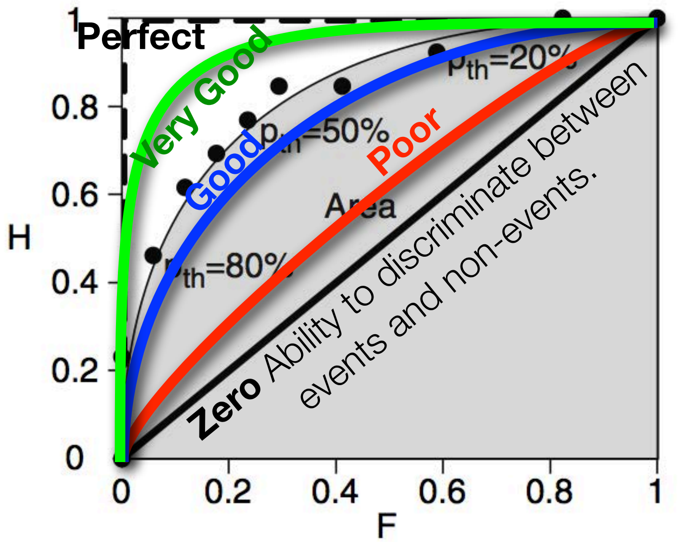


Figure 20.24
ROC diagram plots hit rate (H) vs. false-alarm rate (F) for a range of probability thresholds (p_{th}).

Probability Fcst Verification

ROC Skill Score

Let **A** = area under roc curve (shaded in fig. below)

Define a ROC skill score:

$$SS_{ROC} = (2 \cdot A) - 1$$

SS_{ROC} = 1 for perfect discrimination,
SS_{ROC} =0 for zero discrimination.

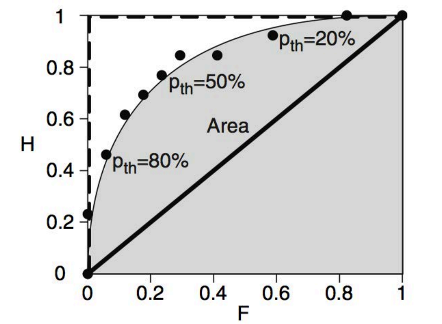
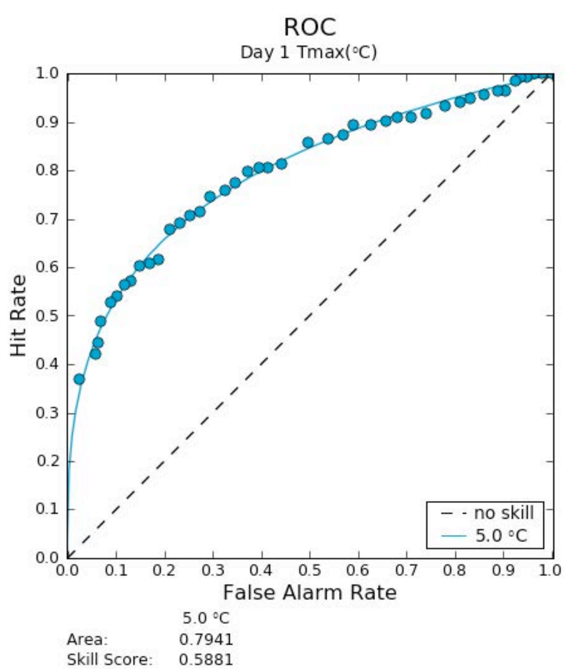


Figure 20.24
ROC diagram plots hit rate (H) vs. false-alarm rate (F) for a range of probability thresholds (p_{th}).

ROC example



ROC example

Change with forecast horizon (days).

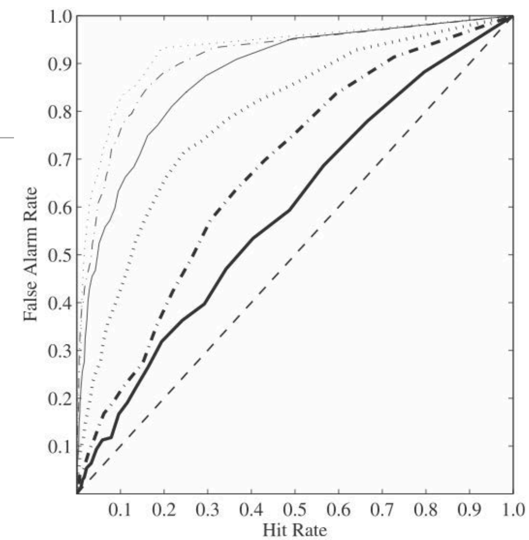
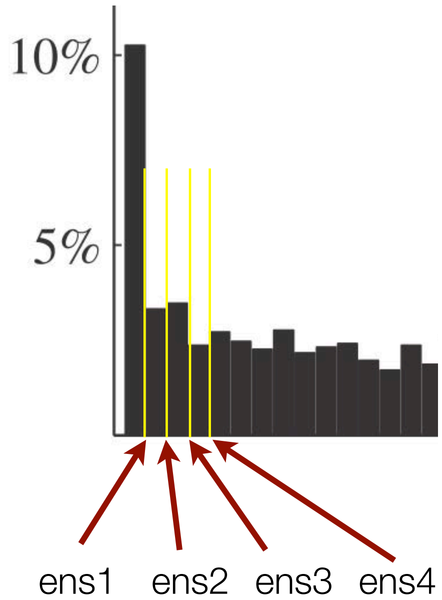


FIG. 6. ROC area plots for a subset of the forecast period for a daily maximum temperature anomaly +5°C threshold. The plots range from an ROC curve for day 2 (dotted line), 4 (dot–dashed line), 6 (solid line), 8 (heavy dotted line), 10 (heavy dot–dashed line); and 12 (heavy solid line). The dashed line is the zero-skill line.

Ensemble Fcst Verification

- **Rank Histogram**
  -or-
  Talagrand diagram

Count number of observations between ensemble members that have been sorted (ranked) from lowest value (of temperature, wind, etc.) to highest.



Ensemble Fcst Verification

A **Flatness Score** for the rank histogram:

$$\Delta = \sum_{i=1}^{M+1} \left(s_i - \frac{N}{M+1} \right)^2 \qquad (3)$$

measures the deviation of the histogram from a horizontal line. For a reliable system, the base value is $\Delta_0 = NM/(M+1)$. The ratio

$$\delta = \Delta/\Delta_0 \qquad (4)$$

where a value closer to 1 is better.
s_i is the count of obs in histogram bar i.

Ensemble Fcst Verification

Interpretation:

shape : ensemble is

- U-shaped: underdispersive
- hill-shaped: overdispersive
- flat : perfect

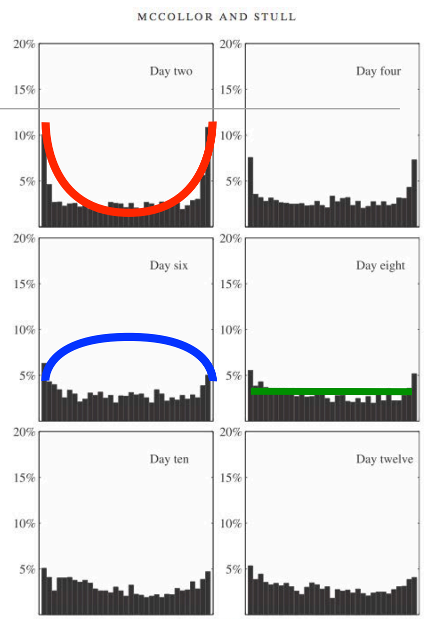


FIG. 10. Rank histograms for the ensemble of daily maximum temperatures for (top to bottom) days 2, 4, 6, 8, 10, and 12.

Ensemble Fcst Verification

The Brier skill score is then expressed as

$$\text{BSS} = \frac{\text{resolution}}{\text{uncertainty}} - \frac{\text{reliability}}{\text{uncertainty}}$$

$$= \text{relative resolution} - \text{relative reliability}.$$

A perfect forecast would have the relative resolution equal to 1 and the relative reliability equal to 0.

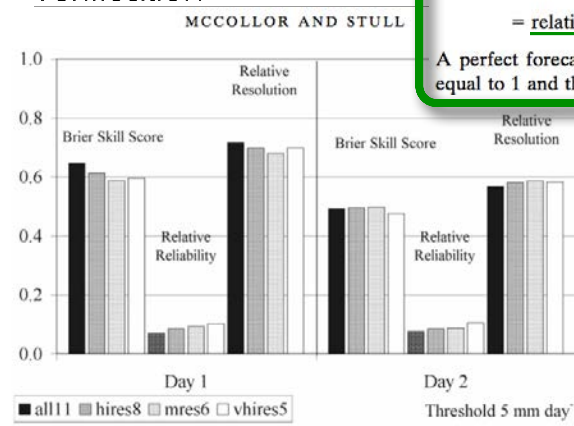


FIG. 8. BSS (1 is perfect), relative reliability (0 is perfect), and relative resolution (1 is perfect) for the SREFs for forecast days 1 and 2. The precipitation threshold is 5 mm day^{-1}. The configuration encompassing all 11 ensemble members (from 108 km down to 4 km for all three models) performs better for the day-1 forecast than do the other configurations with fewer ensemble members. The four SREF configurations show similar Brier skill scores for the day-2 forecast. Note that resolution deteriorates in the day-2 time frame while reliability is consistent for both the day-1 and day-2 forecasts.

Ensemble Fcst Verification

Sharpness

Associated with reliability is sharpness, which characterizes the relative frequency of occurrence of the forecast probabilities. Sharpness is often depicted in a histogram indicating the relative occurrence of each forecast probability category. If forecast probabilities are frequently near 0 or near 1, then the forecasts are sharp, indicating the forecasts deviate significantly from the climatological mean, a positive attribute of an ensemble forecast system. Sharpness measures the variability of the forecasts alone, without regard to their corresponding observations; hence, it is not a verification measure in itself. In a perfectly reliable forecast system, sharpness is identical to resolution.

Ensemble Fcst Verification

Resolution

j. Resolution

A useful probabilistic forecast system must be able to a priori differentiate future weather outcomes, so that differing forecasts are, in fact, associated with distinct verifying observations. This is the most important attribute of a forecast system (Toth et al. 2003) and is called *resolution*.

Resolution cannot be improved through simple adjustment of probability values or statistical postprocessing. Resolution can be gained only by improving the actual forecast model engine that produces the forecasts.

Ensemble Fcst Verification

Prob. Integral Transform (PIT)

Let $f_t(x)$ = **forecast** prob. density for the value x of any variable (e.g., Temperature)

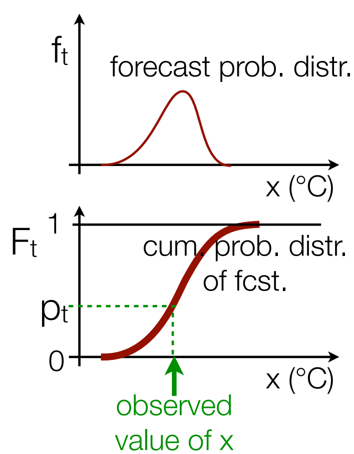
Let $F_t(x)$ denote the forecasted cumulative distribution function (CDF) given by

$$F_t(x) = \int_{-\infty}^{x} f_t(s)\,ds. \qquad (1)$$

In addition, let x_t denote the observed state of X at time t. Let p_t denote the CDF value corresponding to the observed state:

$$p_t = F_t(x_t). \qquad (2)$$

Often, p_t is called the probability integral transform value (PIT value) corresponding to the observation.



Good fcst if $p_t = p$, where p = observed cum. freq at x_t.

Prob. Fcst Verif.

Ignorance Score (IGN)

$$\mathrm{IGN}(f) = \frac{1}{|T|}\sum_{t\in T} -\log_2[f_t(x_t)]. \qquad (21)$$

IGN rewards forecasts that place high confidence in the value where the observation falls. Low ignorance scores are desired.

where $f_t(x_t)$ = **forecast** prob. density for the **observed** value x_t of any variable (e.g., Temperature)

Example of IGN

Top graph shows improvement in ignorance score. larger is better.

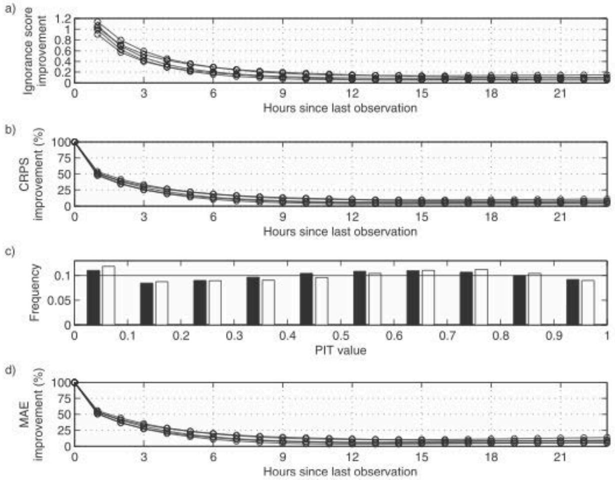
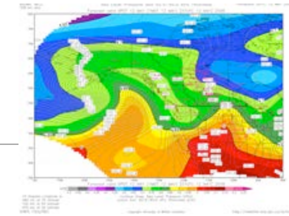


FIG. 5. Verification statistics for the probabilistic forecasts used in the study. (a) Reduction (improvement) of the ignorance score by the updated probabilistic forecast relative to the original probabilistic forecast. Each of the five lines represents the score for a different station. (b) Percentage improvement in the CRPS by the updated probabilistic forecast. (c) PIT histogram of the updated forecasts (black bars) and the original forecasts (white bars), indicating the reliability of the forecasts. (d) Percentage improvement in mean absolute error of the median of the updated probability distributions relative to the median of the original distribution.

Verification Topics



- Continuous Variables [T, U, V, P, bounded (RH, Precip)...]
- Categorical (Binary, yes/no) Events
- Probabilistic & Ensemble Forecasts
- Terrain Issues

Terrain Issues

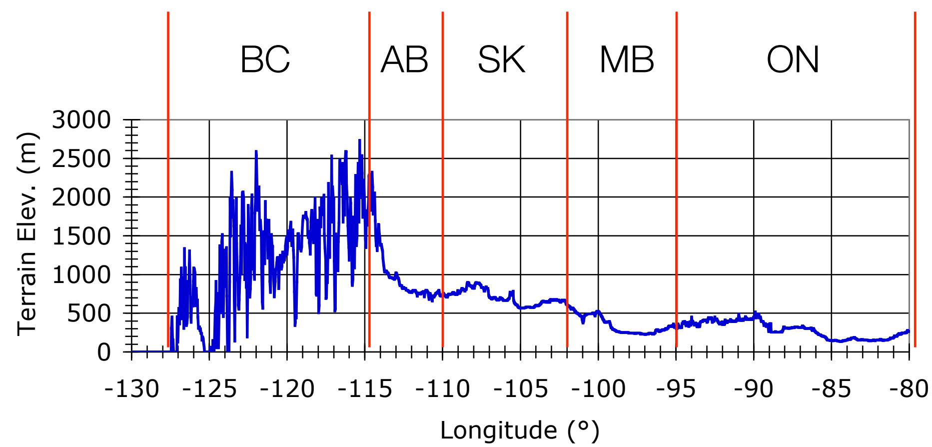
Modeled terrain is smoothed relative to actual terrain.

Mountain tops are cut off, and valleys are filled in. More so for coarser grids.

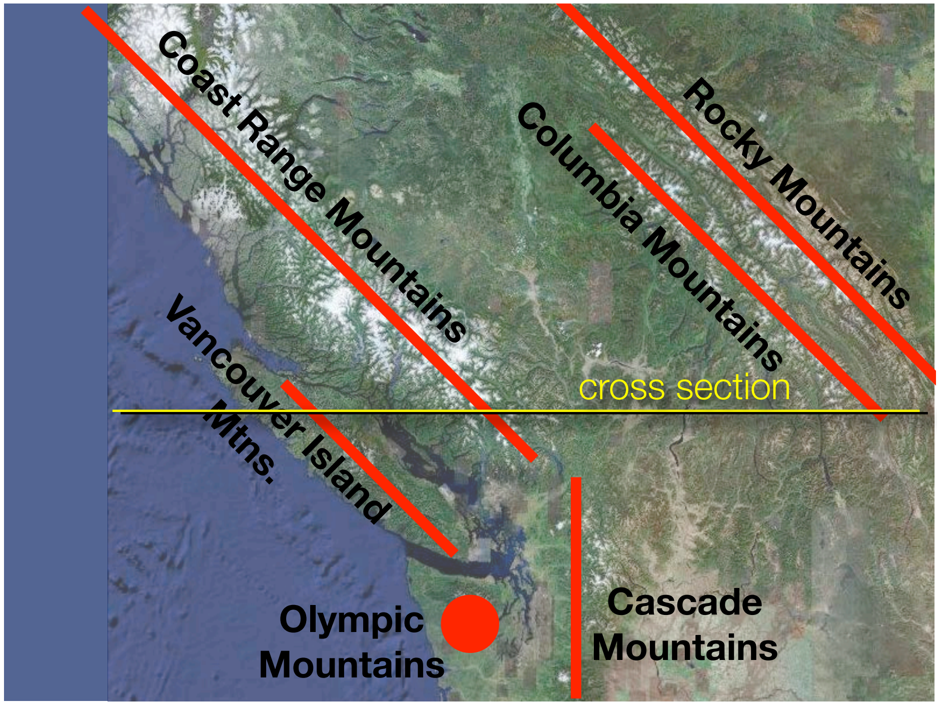
Thus, the elevation of a verifying obs is often different from the elevation in the NWP model.

Thus, the model is **not representative** of the observation.

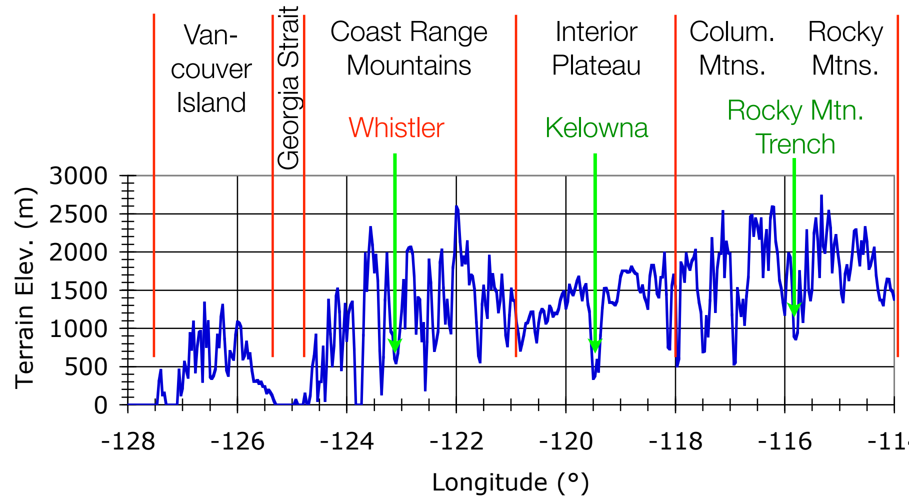
Canadian Terrain Elevation



- West-East terrain cross section through Whistler (50.12°N)

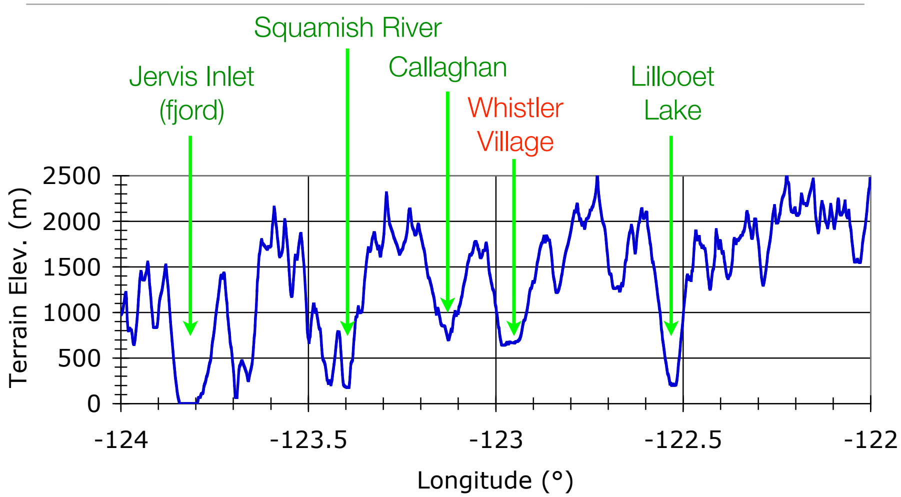


British Columbia Terrain Elevation



- West-East terrain cross section through Whistler (50.12°N)

Zooming in Near Whistler



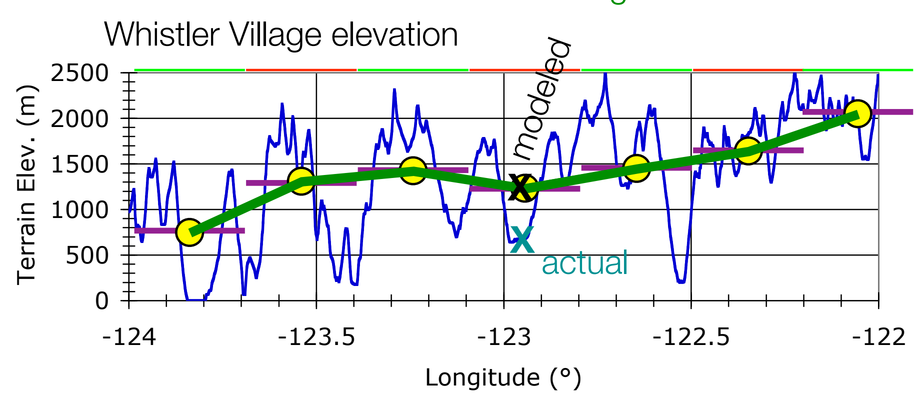
- West-East terrain cross section through Whistler (50.12°N)

Terrain Elevation vs. Grid Size.

Terrain must be smoothed to match grid resolution.

Example: If grid size is Δx = 21 km

Then smoothed terrain is shown in green.



- West-East terrain cross section through Whistler (50.12°N), where 0.1°lon ≈ 7 km.

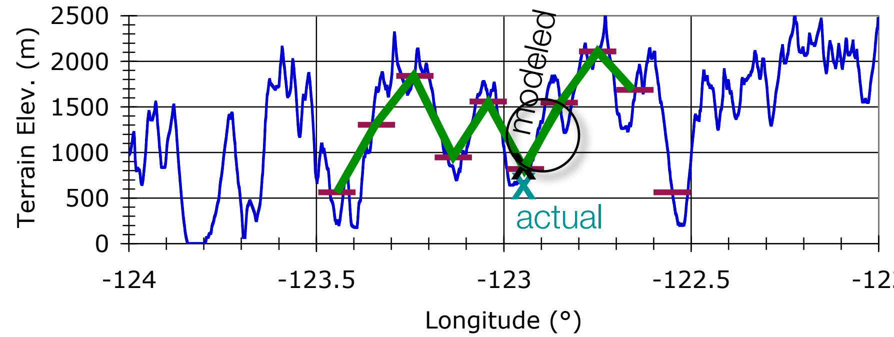
How Fine is Fine Enough?
Many valleys are narrower than 1 km

An Obvious Trick: Use finer horizontal grid size

Example: If grid size is $\Delta x = 7$ km

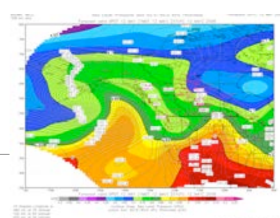
Then the modeled terrain is closer to the actual terrain. **Good.**

And the modeled slopes become steeper (closer to real). **Difficult**.



- West-East terrain cross section through Whistler (50.12°N), where 0.1°lon ≈ 7 km.

Summary



- Continuous Variables [T, U, V, P, bounded (RH, Precip)...]
- Categorical (Binary, yes/no) Events
- Probabilistic & Ensemble Forecasts
- Terrain Issues

Roland Stull, UBC