TECLIM internal lectures

10 March 2022

Verification of uncertain forecasts

François Massonnet

francois.massonnet@uclouvain.be

@FMassonnet

UCLouvain

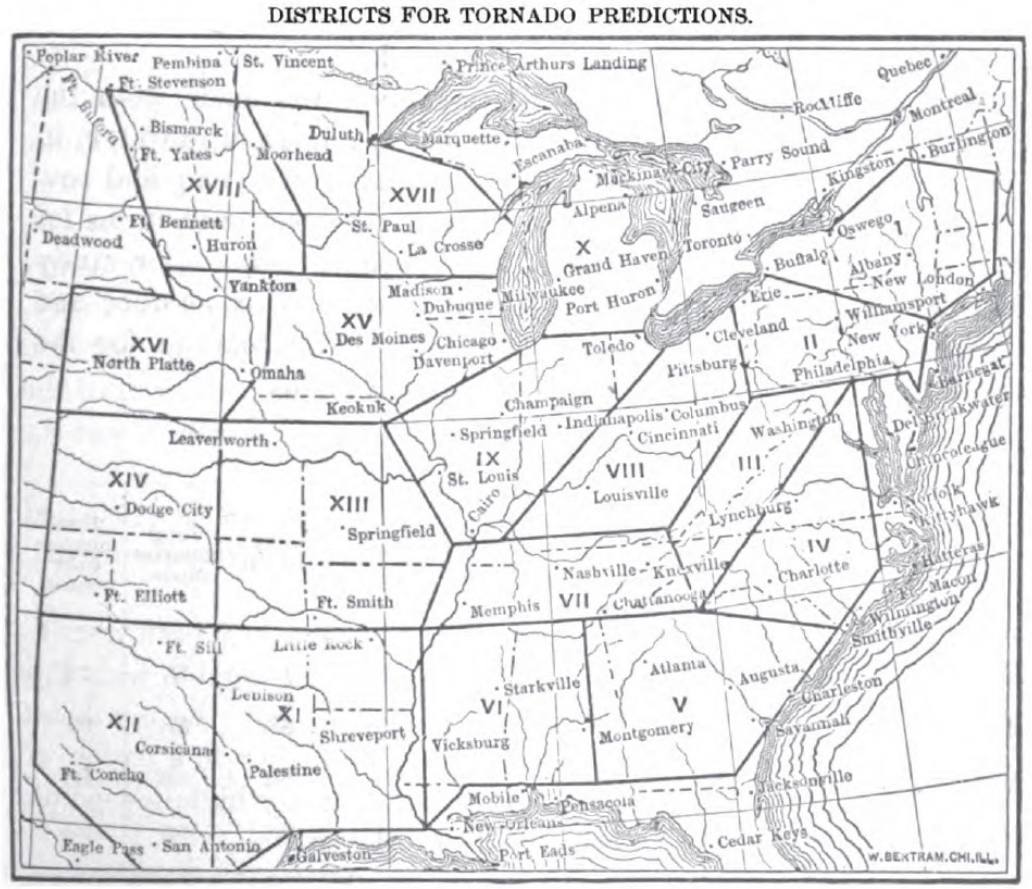
fnrs LA LIBERTÉ DE CHERCHER

EARTH & LIFE INSTITUTE

1. Historical perspective
2. Forecast verification defined
3. Will you go swimming this summer?
4. Scoring rules, scores and skill scores
5. Verification of full pdf forecasts
6. Visual representation of skill
7. Everything that was not said…

1. Historical perspective
2. Forecast verification defined
3. Will you go swimming this summer?
4. Scoring rules, scores and skill scores
5. Verification of full pdf forecasts
6. Visual representation of skill
7. Everything that was not said…

The « Finley affair » (1884)



Funnel-shaped cloud or tornado

Finley, P. (1884). Tornado Predictions. *Amer. Meteor. J.*, *1*, 85–88.

The « Finley affair » (1884)

Percentage of Tornado Predictions.

DISTRICTS.	I.	II.	III.	IV.	V.	VI.	VII.	VIII.	IX.	X.	XI.	XII.	XIII.	XIV.	XV.	XVI.	XVII.	XVIII.	General Average Percentage.
March..... Eight hour predictions.	100.00	98.86	95.46	88.53	95.92	92.56	88.53	87.86	89.10	100.00	98.80	100.00	91.32	100.00	99.43	100.00	100.00	100.00	95.61
April .. .. Eight hour predictions.	100.00	100.00	100.00	95.28	96.22	95.28	100.00	98.11	99.04	98.11	99.01	100.00	94.07	100.00	98.08	100.00	100.00	100.00	98.51
May..... Eight hour predictions.	100.00	100.00	100.00	100.00	100.00	100.00	100.00	96.77	96.77	96.77	100.00	100.00	95.16	96.77	96.77	96.77	100.00	100.00	98.65
May..... Sixteen hour predictions.	100.00	100.00	100.00	95.16	98.39	98.39	96.77	90.32	88.71	96.77	98.10	100.00	82.50	100.00	97.67	100.00	100.00	100.00	96.54

Finley, P. (1884). Tornado Predictions. *Amer. Meteor. J.*, *1*, 85–88.

The « Finley affair » (1884)

	Observations		
Forecasts	Tornado	No tornado	
Tornado	28 *a*	72 *b*	100
No tornado	23 *c*	2680 *d*	2703
	51	2752	2803

Contingency table of all Finley's forecasts (from Murphy, 1996)

- Finley used the « Proportion correct » metric:

$$PC = \frac{a+d}{a+b+c+d} = 96.6\%$$

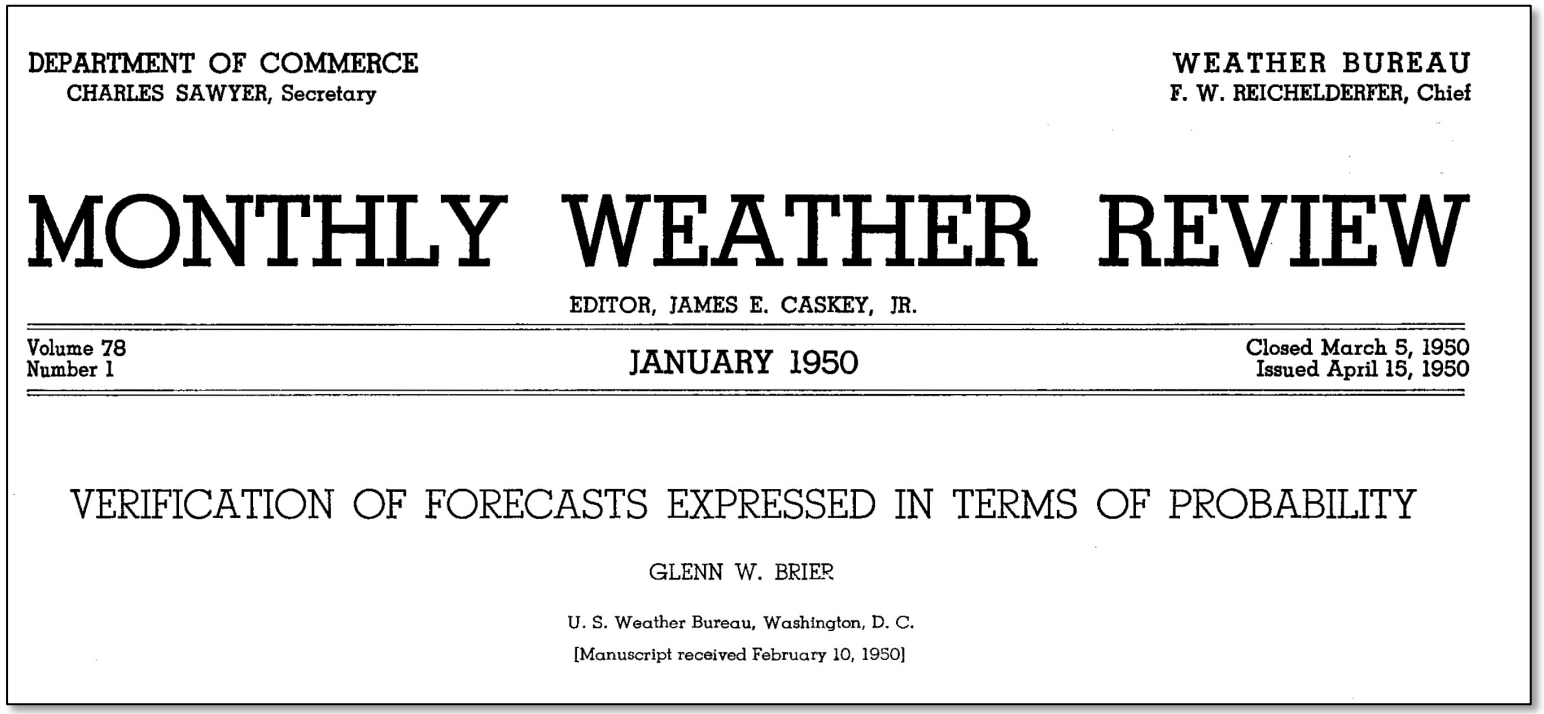
- Gilbert (1884) challenged Finley's method

$$PC = \frac{0+2752}{2803} = 98.1\% > 96.6\%$$

Murphy, A. H. (1996). The Finley Affair: A Signal Event in the History of Forecast Verification. *Weather and Forecasting*, *11*(1), 3–20. https://doi.org/10.1175/1520-0434(1996)011<0003:TFAASE>2.0.CO;2

Gilbert, G. K. (1884). Finley's Tornado Predictions. *Amer. Meteor. J.*, *1*, 166–172.

Accounting for uncertainty in forecast verification (1950)



DEPARTMENT OF COMMERCE
CHARLES SAWYER, Secretary

WEATHER BUREAU
F. W. REICHELDERFER, Chief

MONTHLY WEATHER REVIEW

EDITOR, JAMES E. CASKEY, JR.

Volume 78
Number 1

JANUARY 1950

Closed March 5, 1950
Issued April 15, 1950

VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY

GLENN W. BRIER

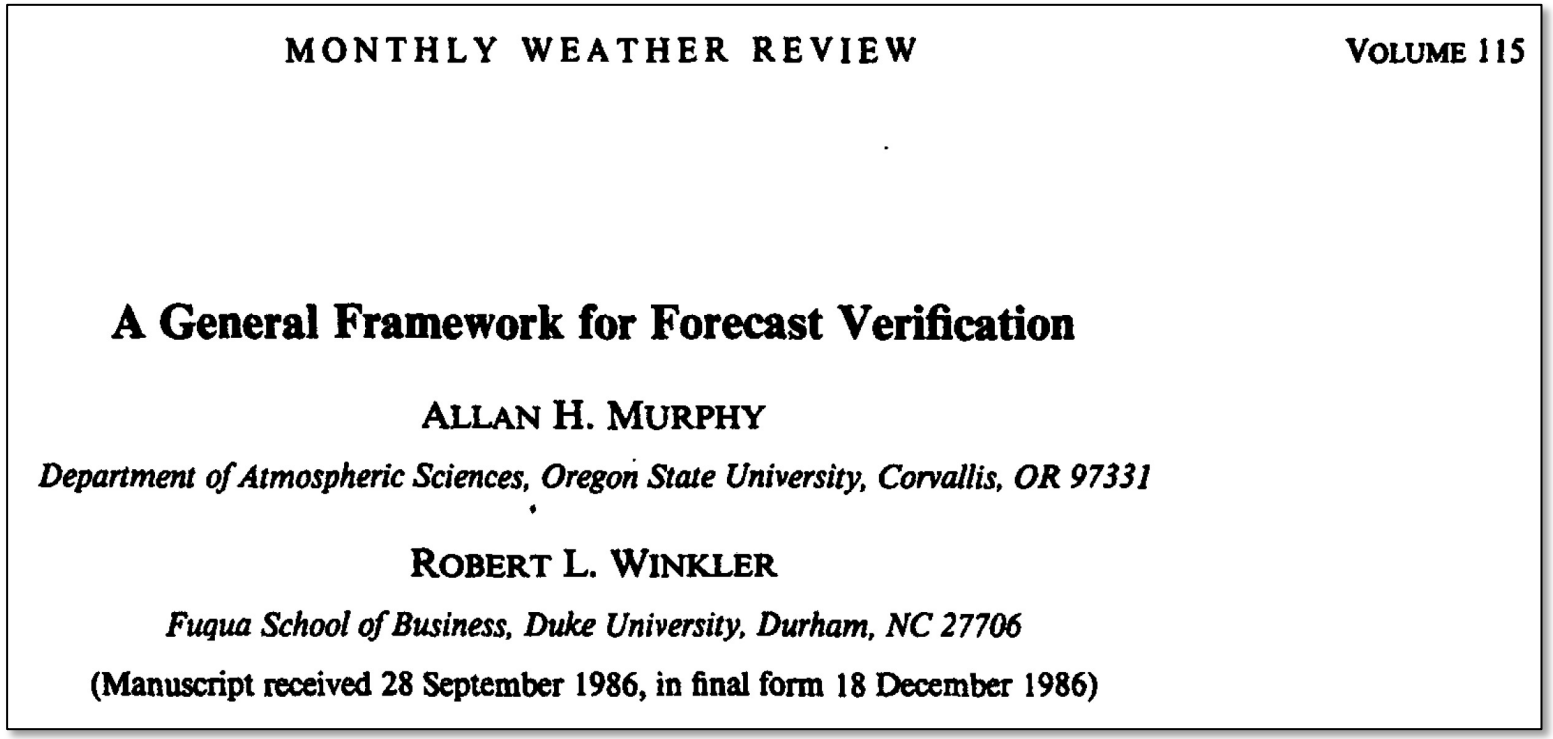
U. S. Weather Bureau, Washington, D. C.

[Manuscript received February 10, 1950]

« It is the purpose of this paper to discuss one situation where it appears to be possible to devise a verification scheme that cannot influence the forecaster in any undesirable way »

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

A theoretical background for forecast verification (1987)



MONTHLY WEATHER REVIEW VOLUME 115

A General Framework for Forecast Verification

ALLAN H. MURPHY

Department of Atmospheric Sciences, Oregon State University, Corvallis, OR 97331

ROBERT L. WINKLER

Fuqua School of Business, Duke University, Durham, NC 27706

(Manuscript received 28 September 1986, in final form 18 December 1986)

Murphy, A. H., & Winkler, R. L. (1987). A General Framework for Forecast Verification. *Monthly Weather Review*, *115*(7), 1330–1338. https://doi.org/10.1175/1520-0493(1987)115<1330:AGFFFV>2.0.CO;2

1. Historical perspective
2. Forecast verification defined
3. Will you go swimming this summer?
4. Scoring rules, scores and skill scores
5. Verification of full pdf forecasts
6. Visual representation of skill
7. Everything that was not said…

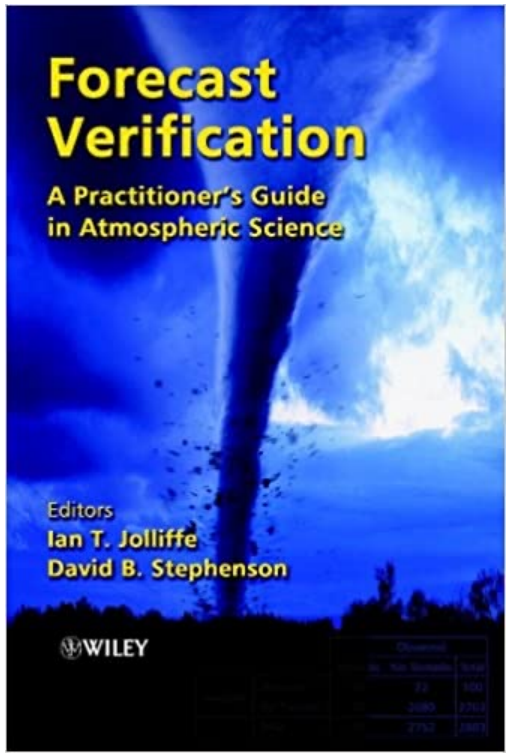
1. Historical perspective
2. Forecast verification defined
3. Will you go swimming this summer?
4. Scoring rules, scores and skill scores
5. Verification of full pdf forecasts
6. Visual representation of skill
7. Everything that was not said…

Definition « with the hands »

« Forecasts are almost always made and used in the belief that having a forecast available is **preferable** to remaining in complete **ignorance** about the future event of interest.

It is important to **test this belief a posteriori** by assessing **how skilful or valuable** was the forecast.

This is the topic of **forecast verifcation**. »



Jolliffe, I. T., & Stephenson, D. B. (Eds.). (2003). *Forecast verification: A practitioner's guide in atmospheric science*. J. Wiley.

A more formal definition

Forecast verification is the attempt to characterize the probability density function (pdf) of a set of forecasts and verifying observations...

$$p(f, o)$$

Forecasts

Verifying observations

... given a finite sample of forecast-observation pairs

$$\{f_t, o_t\} \quad t = 1, \ldots, T$$

Murphy, A. H., & Winkler, R. L. (1987). A General Framework for Forecast Verification. *Monthly Weather Review*, *115*(7), 1330–1338. https://doi.org/10.1175/1520-0493(1987)115<1330:AGFFFV>2.0.CO;2

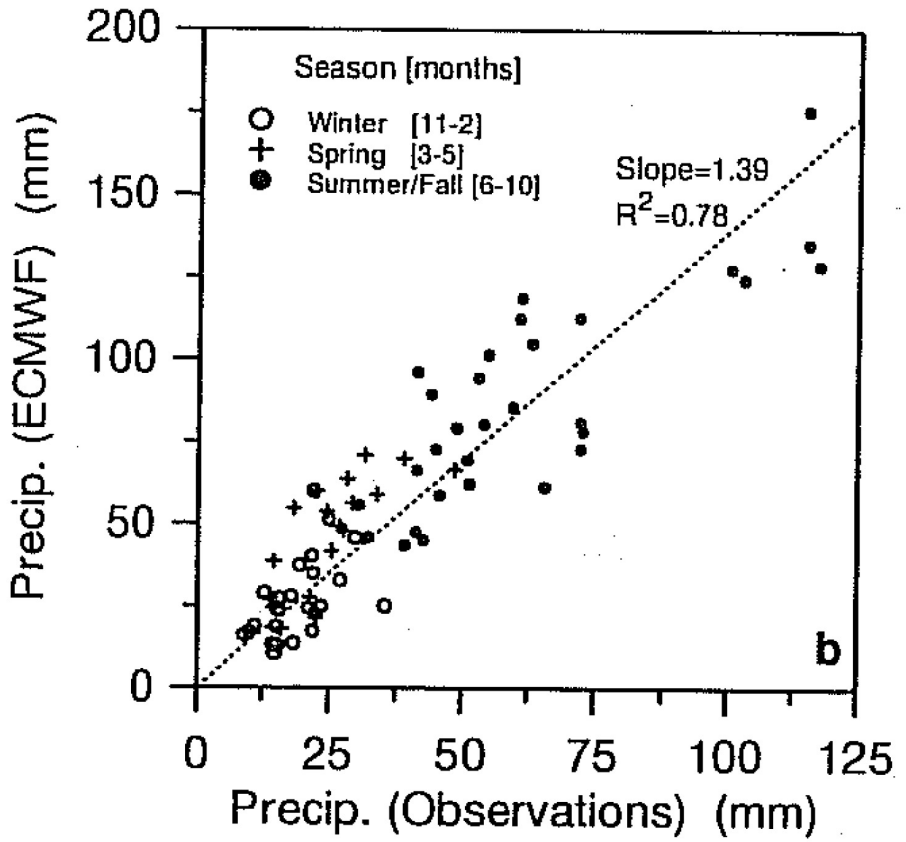
Four types of forecasts

	Deterministic	Probabilistic
Value-based	*Single-value forecasts*	*Full pdf forecasts*
Event-based	*Yes/No forecasts*	*Categorical probabilistic forecasts*

Four types of forecasts

	Deterministic	Probabilistic
Value-based	*Single-value forecasts*	*Full pdf forecasts*
Event-based	*Yes/No forecasts*	*Categorical probabilistic forecasts*

Single-value forecasts



$$(f, o) \in \mathbb{R}^2$$

$$p(f, o)$$

↓

2-D cloud of points

↓

- Bias
- R-squared
- Pearson / Spearman Correlation
- (Root) Mean Squared Error
- Mean absolute error

…

Betts, P. Viterbo, Anton Beljaars, & B.J.J. van den Hurk. (1999). Use of field data to diagnose land-surface interaction. *Diagnosis of Models and Data Assimilation Systems, 6-10 September 1999*. Diagnosis of models and data assimilation systems, 6-10 September 1999, Shinfield Park, Reading. https://www.ecmwf.int/node/8221

Four types of forecasts

	Deterministic	Probabilistic
Value-based	*Single-value forecasts*	*Full pdf forecasts*
Event-based	*Yes/No forecasts*	*Categorical probabilistic forecasts*

Four types of forecasts

	Deterministic	Probabilistic
Value-based	*Single-value forecasts*	*Full pdf forecasts*
Event-based	*Yes/No forecasts*	*Categorical probabilistic forecasts*

Yes/No forecasts

$$(f, o) \in \{0, 1\}^2$$

(0 = "no", 1 = "yes")

$$p(f, o)$$

Forecasts	Observations: Tornado	Observations: No tornado	
Tornado	28	72	100
No tornado	23	2680	2703
	51	2752	2803

Joint distribution

$p(f, o)$

p(f = yes, o = yes) = 28/2803 = 1.0 %
p(f = yes, o = no) =72/2803 = 2.6%
p(f = no, o = yes) = 23/2803 = 0.8%
p(f = no, o = no) = 2680/2752 = 95.6%

→ p(f = yes) = 100 / 2803 = 3.6 %
→ p(o = yes) = 51 / 2752 = 1.9 %

→ Marginal distributions

$p(o)$
$p(f)$

→ p(f = yes | o = yes) = 28 / 51 = 54.9%
→ p(f = yes | o = no) = 72/ 2752 = 2.6%
→ p(o = yes | f = yes) = 28 / 100 = 28.0%
→ p(o = yes | f = no) = 23 / 2703 = 0.9%

→ Conditional distributions

$p(f|o)$
$p(o|f)$

2 x 2 contingency table summarizing frequencies of occurrence

- Proportion correct
- Hit rate
- False alarm ratio
- Bias
- …

Four types of forecasts

	Deterministic	Probabilistic
Value-based	*Single-value forecasts*	*Full pdf forecasts*
Event-based	*Yes/No forecasts*	*Categorical probabilistic forecasts*

Four types of forecasts

	Deterministic	Probabilistic
Value-based	*Single-value forecasts*	*Full pdf forecasts*
Event-based	*Yes/No forecasts*	*Categorical probabilistic forecasts*

Categorical probabilistic forecasts

$$(f, o) \in [0, 1] \times \{0, 1\}$$

$$p(f, o)$$

↓

2 x N **contingency table** summarizing **per-bin** frequencies of occurrence

Categorical probabilistic forecasts

$p(f, o)$

TABLE 1. Joint and marginal distributions of PoP forecasts and observations for NWS forecaster at Chicago, Illinois.

Joint distribution

		f 0.00	0.02	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	$p(x)$
x	1	0.0014	0.0018	0.0028	0.0138	0.0316	0.0255	0.0223	0.0383	0.0309	0.0426	0.0216	0.0135	0.0032	0.2493
	0	0.0557	0.0500	0.0972	0.1901	0.1773	0.0656	0.0386	0.0337	0.0213	0.0138	0.0074	0.0000	0.0000	0.7507
$p(f)$		0.0571	0.0518	0.1000	0.2039	0.2089	0.0911	0.0609	0.0720	0.0522	0.0564	0.0290	0.0135	0.0032	

→ Marginal distributions

$p(o)$

$p(f)$

	0.00	0.02	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
$p(x=1\mid f)$	0.0245	0.0347	0.0280	0.0677	0.1513	0.2799	0.3662	0.5319	0.5920	0.7553	0.7448	1.0000	1.0000
$p(x=0\mid f)$	0.9755	0.9653	0.9720	0.9323	0.8487	0.7201	0.6338	0.4681	0.4080	0.2447	0.2552	0.0000	0.0000

	f 0.00	0.02	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
$p(f\mid x=1)$	0.0056	0.0072	0.0112	0.0554	0.1268	0.1023	0.0895	0.1536	0.1239	0.1709	0.0866	0.0542	0.0128
$p(f\mid x=0)$	0.0742	0.0666	0.1295	0.2532	0.2362	0.0874	0.0514	0.0449	0.0284	0.0184	0.0099	0.0000	0.0000

→ Conditional distributions

$p(f|o)$

$p(o|f)$

Murphy, A. H., & Winkler, R. L. (1987). A General Framework for Forecast Verification. *Monthly Weather Review*, *115*(7), 1330–1338. https://doi.org/10.1175/1520-0493(1987)115<1330:AGFFFV>2.0.CO;2

Categorical probabilistic forecasts

$$(f, o) \in [0, 1] \times \{0, 1\}$$

$$p(f, o)$$

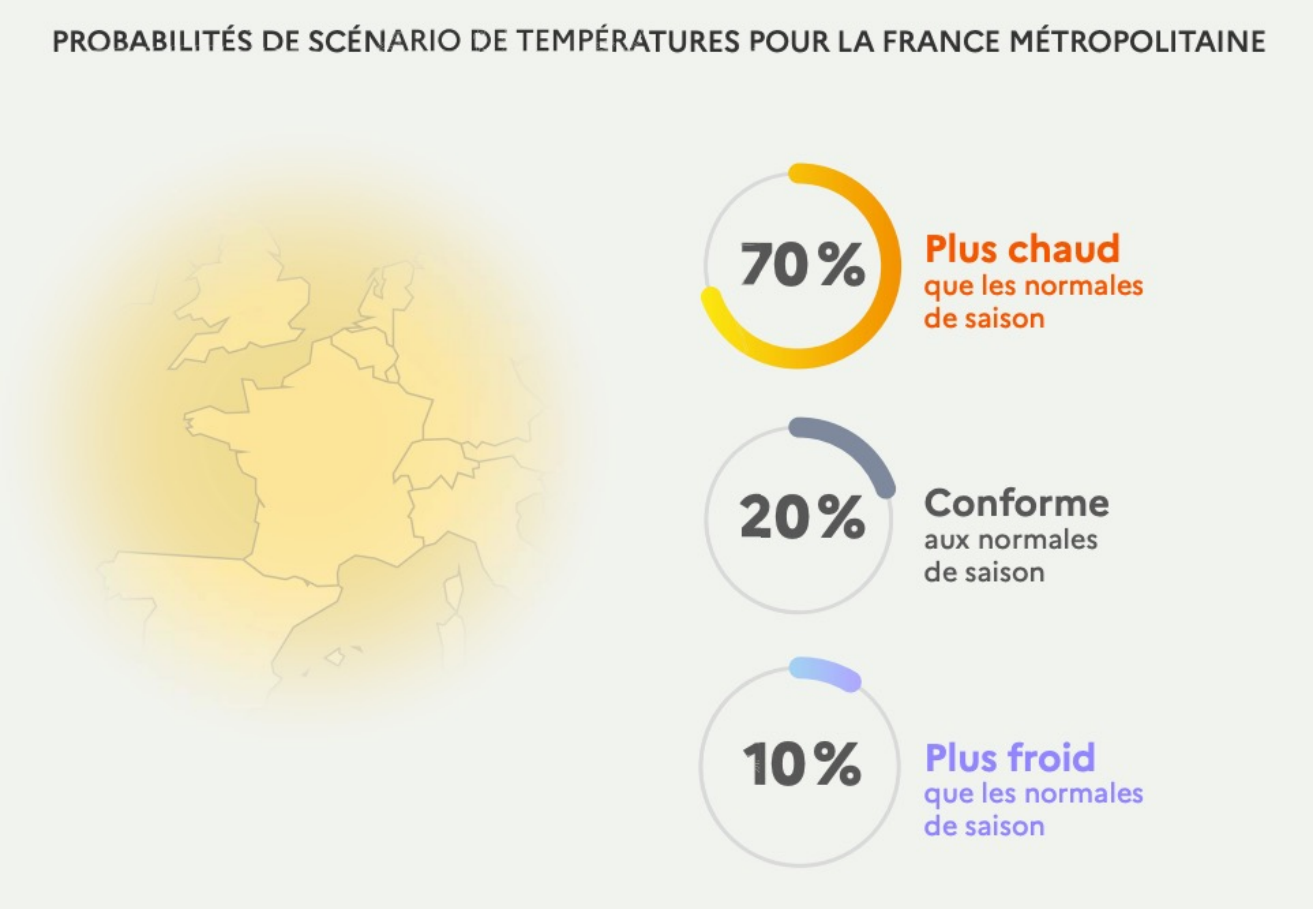
↓

2 x N **contingency table** summarizing **per-bin** frequencies of occurrence

↓

- Brier Score
- Reliability
- Resolution
- Discrimination
- Sharpness
- Ignorance score
- …

Categorical probabilistic forecasts



https://meteofrance.fr/sites/meteofrance.fr/files/files/editorial/Tendances-climatiques-mars-avr-mai2022_V4_0.pdf

Four types of forecasts

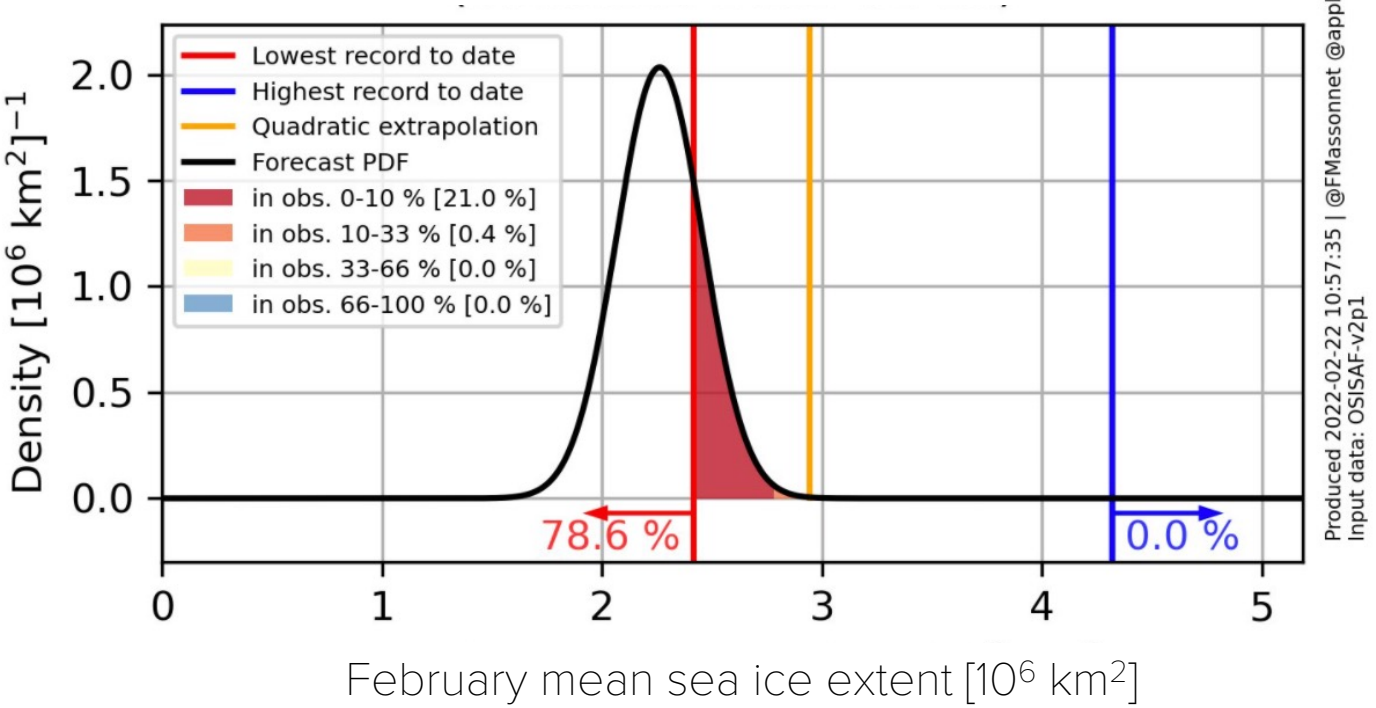
	Deterministic	Probabilistic
Value-based	*Single-value forecasts*	*Full pdf forecasts*
Event-based	*Yes/No forecasts*	*Categorical probabilistic forecasts*

Four types of forecasts

	Deterministic	Probabilistic
Value-based	*Single-value forecasts*	***Full pdf forecasts***
Event-based	*Yes/No forecasts*	*Categorical probabilistic forecasts*

Full pdf forecasts

Forecast pdf of the February 2022 mean Antarctic sea ice extent (issued 22 Feb 2022)



https://twitter.com/FMassonnet/status/1496072560810856451

http://www.climate.be/users/fmasson/sea-ice-forecasts/oper.html

$$(f, o) \in \text{pdf} \times \mathbb{R}$$

$$p(f, o)$$

↓

Difficult to visualize

↓

- Continuous Rank Probability Score
- Rank histogram
- …

Four types of forecasts

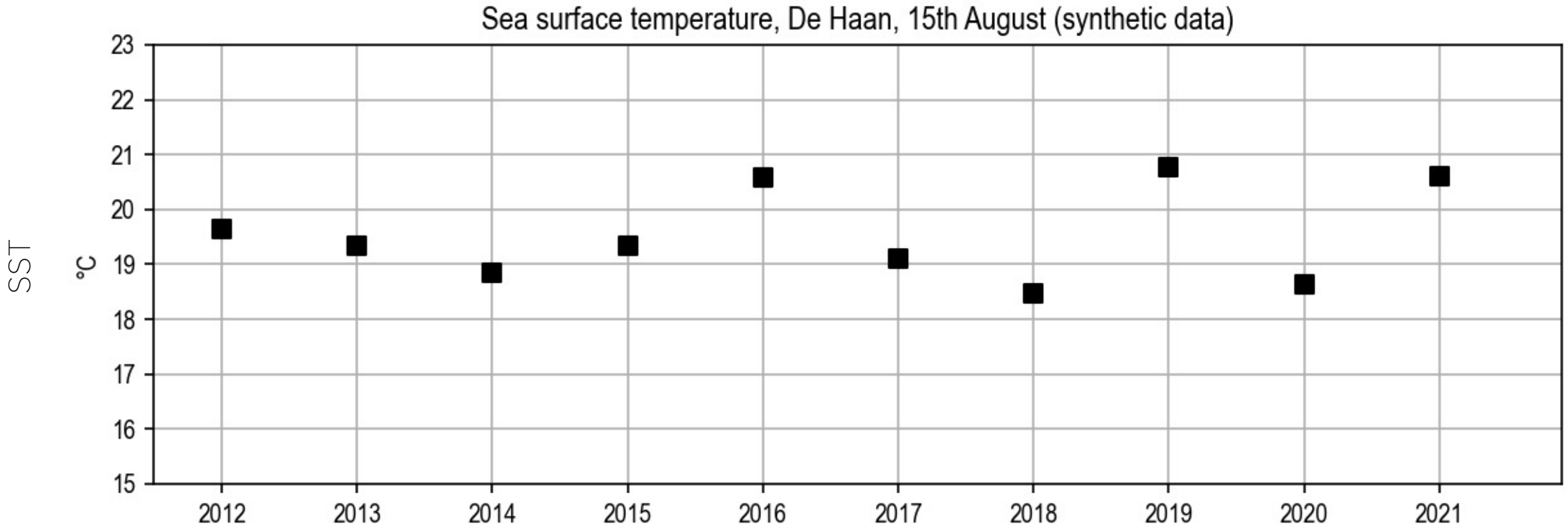
	Deterministic	Probabilistic
Value-based	*Single-value forecasts*	*Full pdf forecasts*
Event-based	*Yes/No forecasts*	*Categorical probabilistic forecasts*

Four types of forecasts

	Deterministic	Probabilistic
Value-based	*Single-value forecasts*	*Full pdf forecasts*
Event-based	*Yes/No forecasts*	*Categorical probabilistic forecasts*

1. Historical perspective
2. Forecast verification defined
3. Will you go swimming this summer?
4. Scoring rules, scores and skill scores
5. Verification of full pdf forecasts
6. Visual representation of skill
7. Everything that was not said…

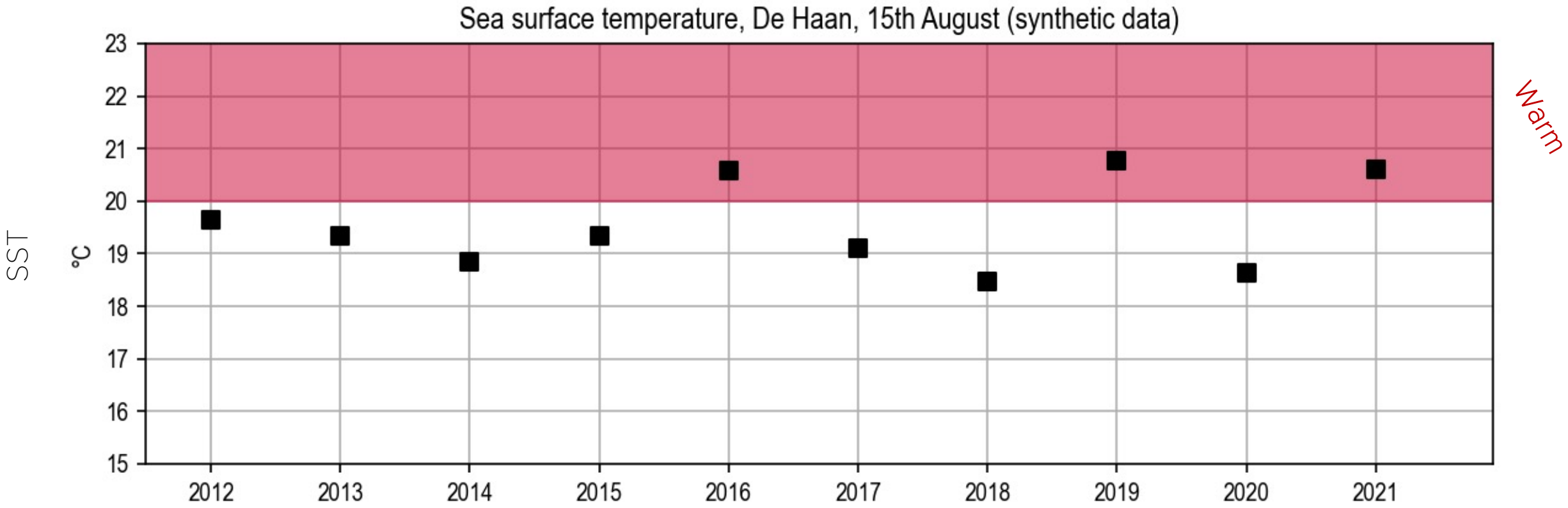
1. Historical perspective
2. Forecast verification defined
3. Will you go swimming this summer?
4. Scoring rules, scores and skill scores
5. Verification of full pdf forecasts
6. Visual representation of skill
7. Everything that was not said…

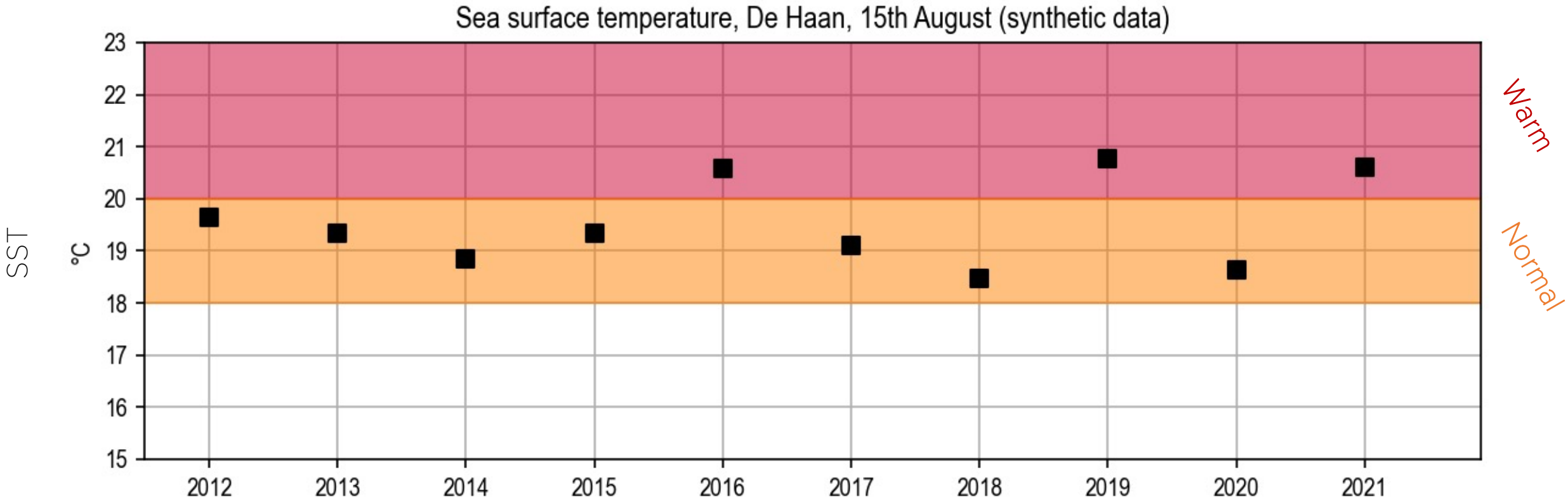


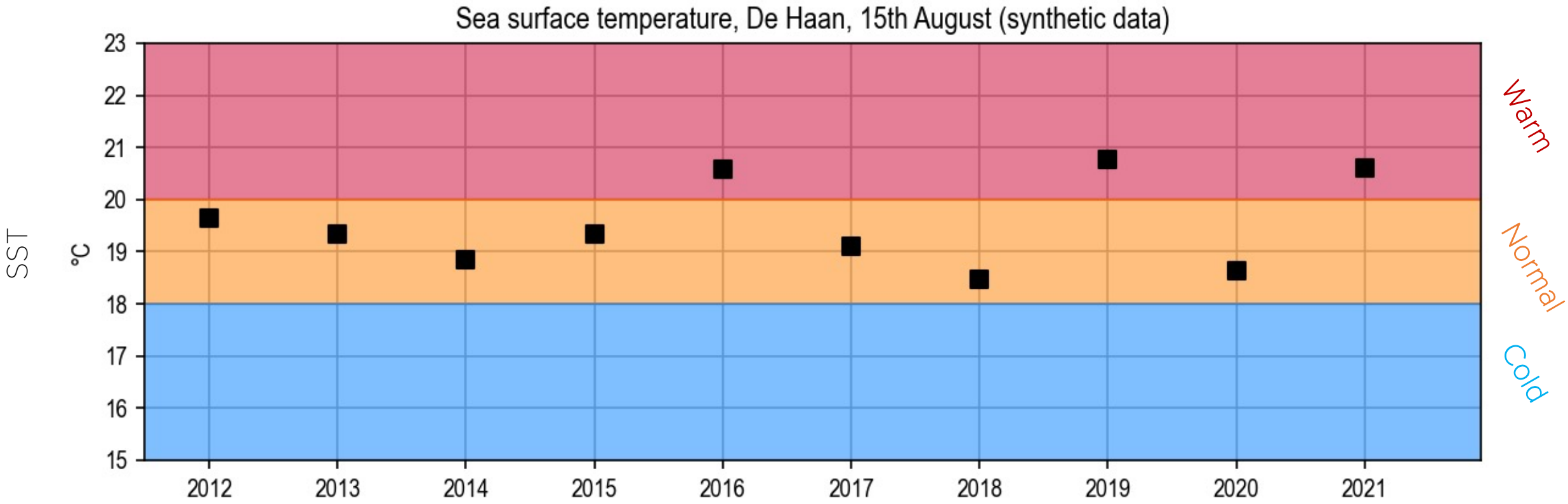
$$o_t \sim \mathcal{N}(\mu = 19\ ^\circ\mathrm{C}, \sigma^2 = 1.0\ ^\circ\mathrm{C}^2)$$

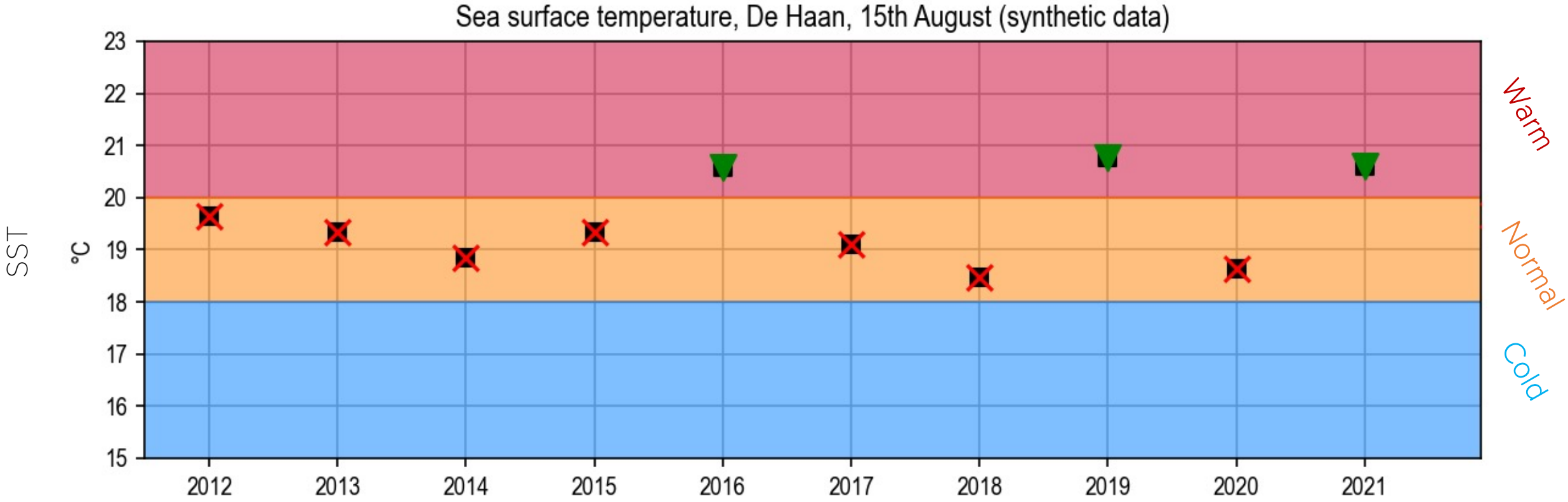
$$t = 1, 2, \ldots, 2021$$

https://github.com/fmassonn/teaching/tree/master/OtherLectures/lectureVerification


Sea surface temperature, De Haan, 15th August (synthetic data)
SST
°C
23
22
21
20
19
18
17
16
15
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
Warm

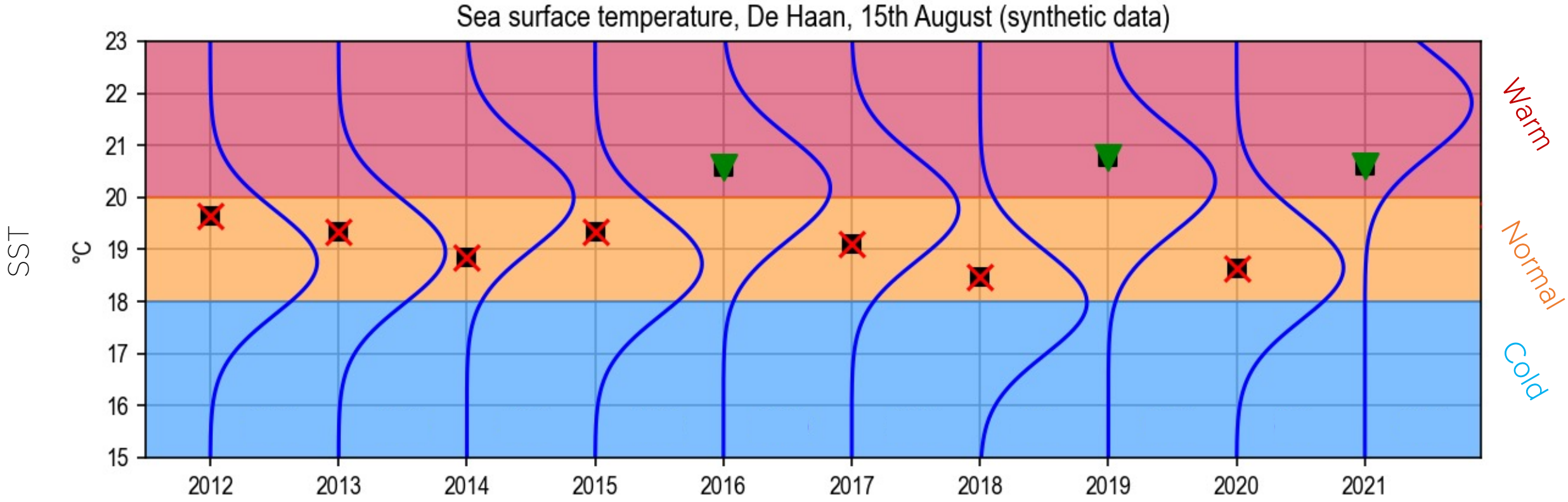

Sea surface temperature, De Haan, 15th August (synthetic data)
SST
°C
23
22
21
20
19
18
17
16
15
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
Warm
Normal


Sea surface temperature, De Haan, 15th August (synthetic data)
SST
°C
23
22
21
20
19
18
17
16
15
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
Warm
Normal
Cold



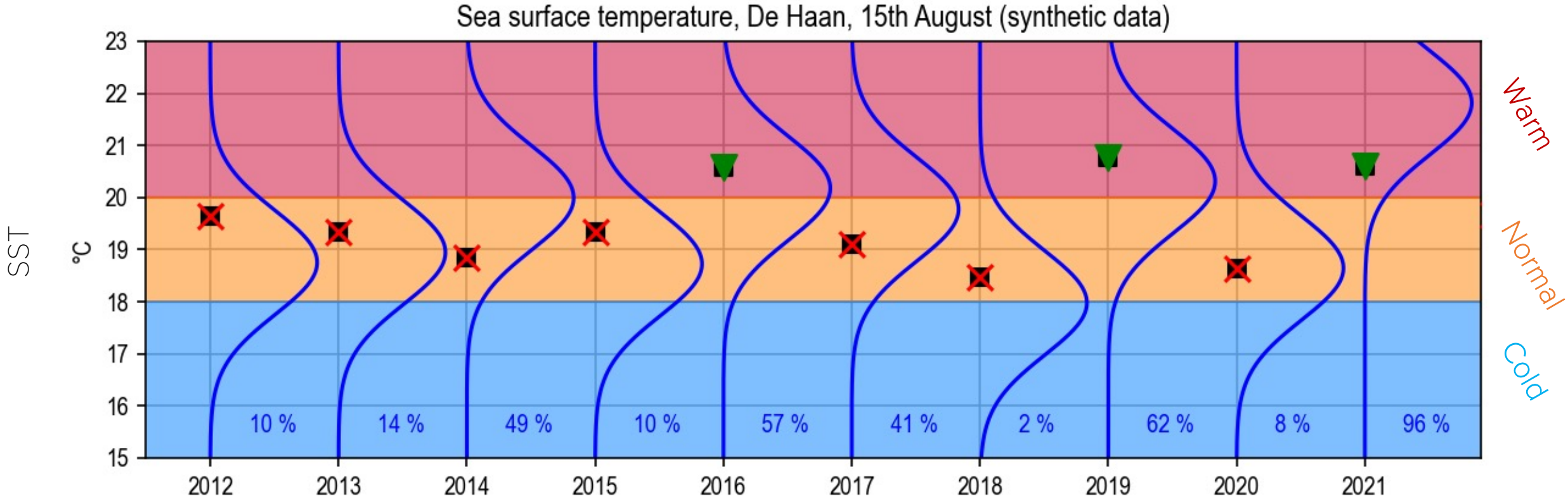
$$E : \mathrm{SST} \geq 20°\mathrm{C}$$

Baseline (climatological) rate of occurence of the event: 15%



Alice

$$f_t^A = \mathcal{N}(o_t + \mathcal{N}(0, (0.5°\mathrm{C})^2), \sigma^2)$$



$p(f^A, o)$ ⟵ $p(f^A)$ ⟶

	0-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%	80-90%	90-100%	sum
$o = 1$	0%	0%	1%	1%	1%	2%	3%	3%	2%	2%	15%
$o = 0$	41%	16%	10%	8%	5%	3%	2%	1%	0%	0%	85%
sum	41%	16%	11%	9%	6%	5%	5%	3%	3%	2%	

Alice

$$f_t^A = \mathcal{N}(o_t + \mathcal{N}(0, (0.5°\mathrm{C})^2), \sigma^2)$$

« Calibration-refinement factorization »

$$p(f, o) = p(o|f)p(f)$$

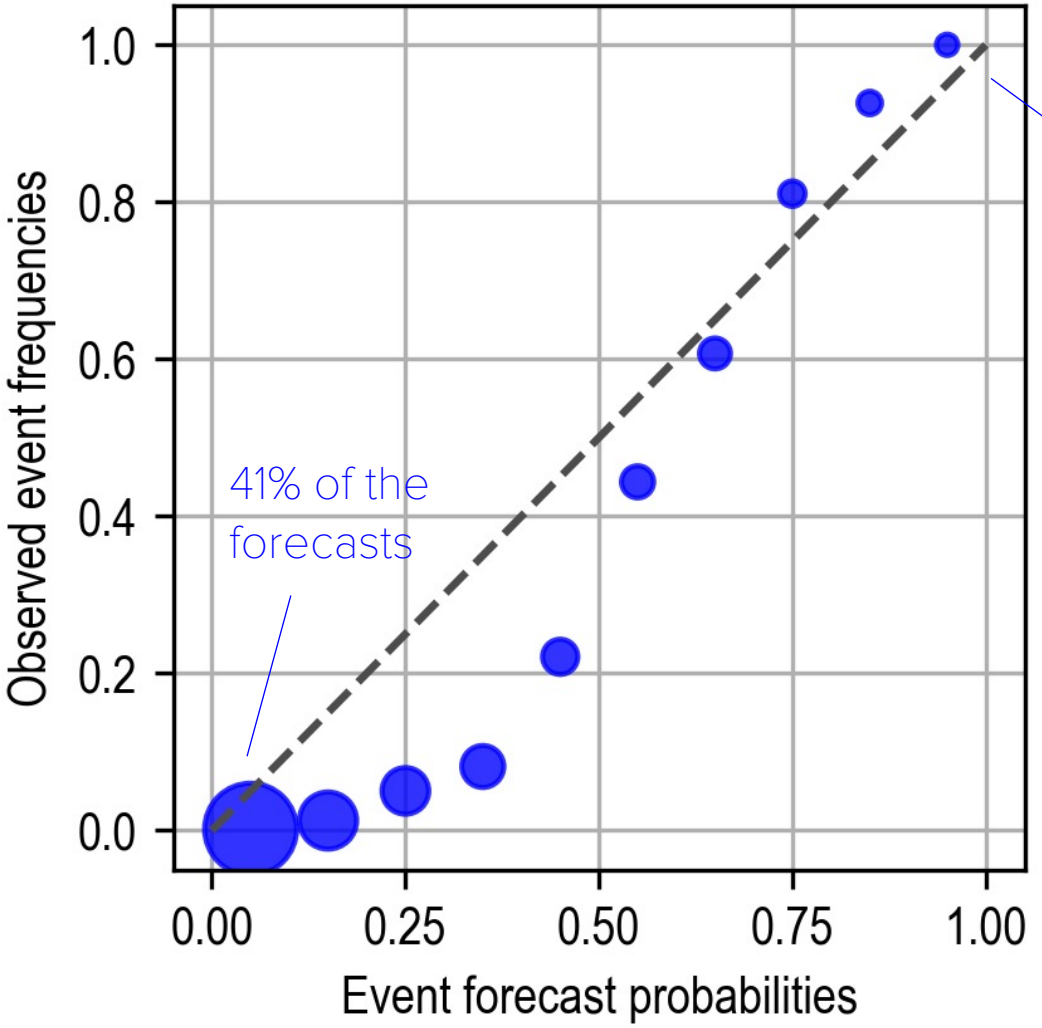
For each probability class $f^i \quad i = 1, \ldots, I$

1. Sub-select all cases where the issued forecast probability equaled f^i
2. Report the observed frequency of occurrence of the event for those cases
3. Ask whether the observed frequency match the forecast probability?

« Reliability »

$$p(o = 1|f = f_i) \stackrel{?}{=} f_i$$

Reliability diagram of Alice



(Bubble size is proportional to subsample size)

Reliability of probabilistic forecasts

$$p(f, o) = p(o|f)p(f)$$

A forecast system is said to be **reliable** (or calibrated) if it provides unbiased estimates of the observed frequencies associated with **each possible forecast probability**

$$p(o \mid f = f_i) = f_i \quad \forall 0 \leq f_i \leq 1$$

1 For each possible predicted probability class (event happens with 0, 1, 2, 3, … 100 % chance)…

2 … the frequency distribution of observations…

3 … that follow the forecast with the said probability…

4 …equals that probability

Murphy, A. H., & Winker, R. L. (1987). A General Framework for Forecast Verification. *Monthly Weather Review*, *115*(7), 1330–1338. https://doi.org/10.1175/1520-0493(1987)115<1330:AGFFFV>2.0.CO;2

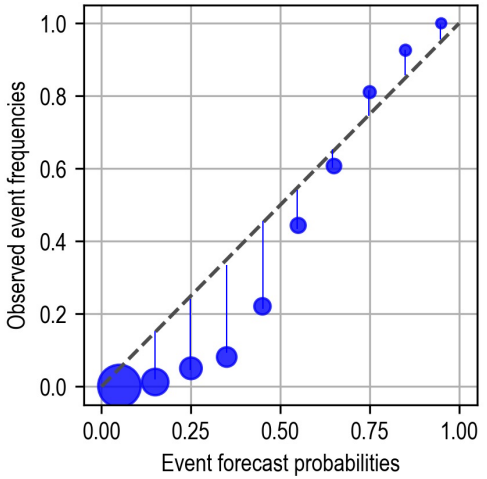
Reliability of probabilistic forecasts

$$p(f, o) = p(o|f)p(f)$$

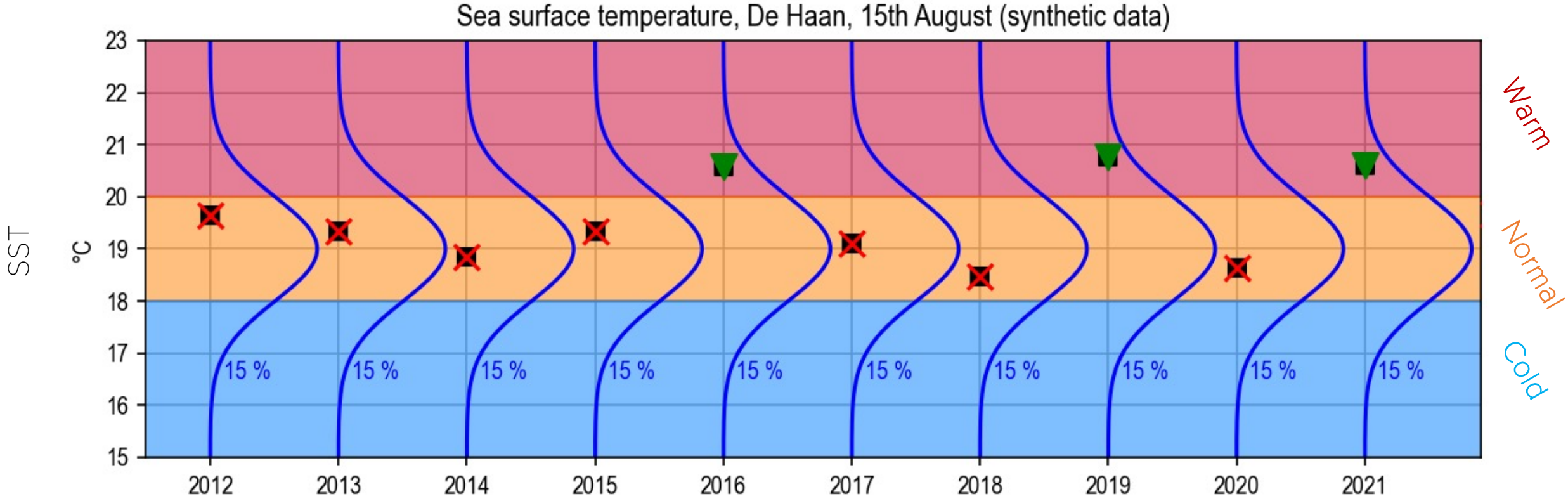
A forecast system is said to be **reliable** (or calibrated) if it provides unbiased estimates of the observed frequencies associated with **each possible forecast probability**

Reliability

- is a property of the forecasts and the observations
- is affected by the **statistical consistency** between forecast probabilities and observed frequencies
- can be improved by recalibration / post-processing of the forecasts



Murphy, A. H., & Winker, R. L. (1987). A General Framework for Forecast Verification. *Monthly Weather Review*, *115*(7), 1330–1338. https://doi.org/10.1175/1520-0493(1987)115<1330:AGFFFV>2.0.CO;2



$p(f^B)$

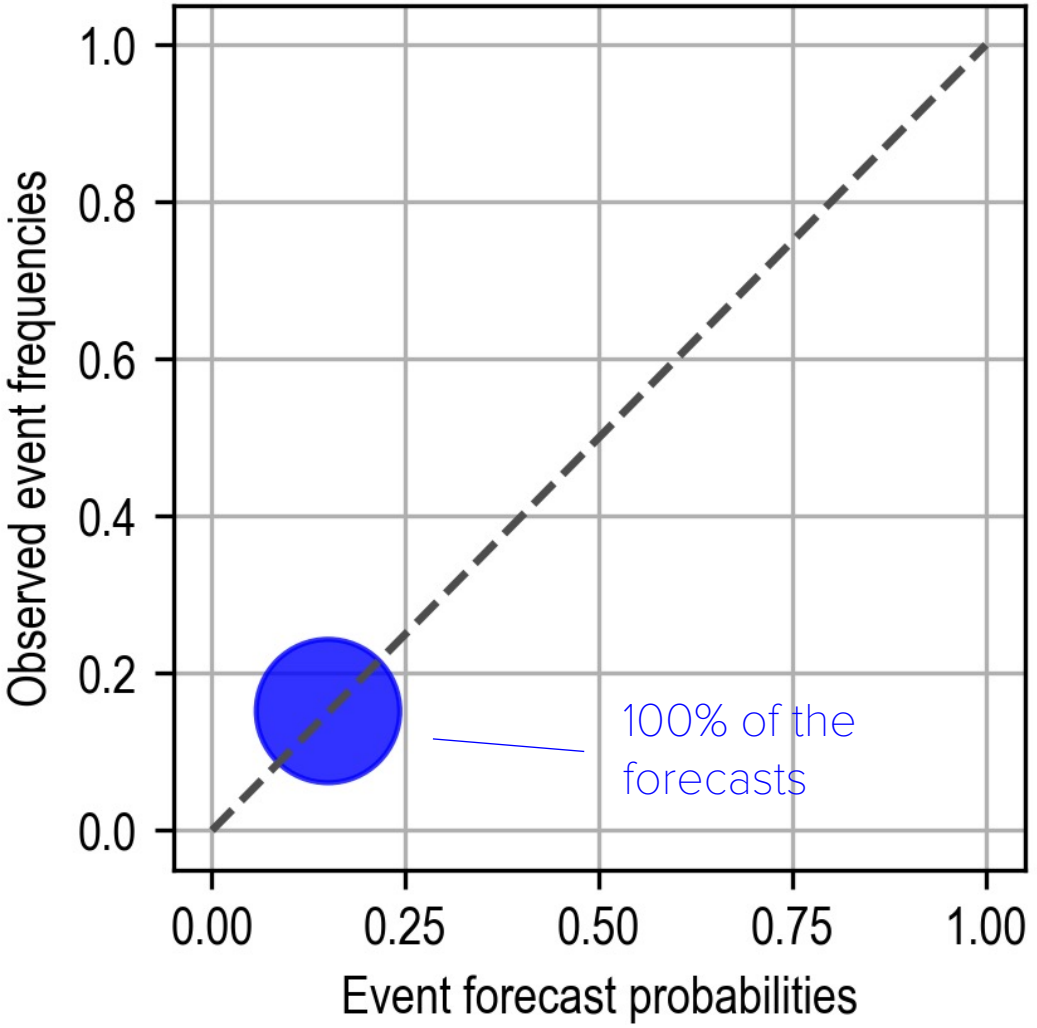
	0-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%	80-90%	90-100%	
$o = 1$	0%	15%	0%	0%	0%	0%	0%	0%	0%	0%	15%
$o = 0$	0%	85%	0%	0%	0%	0%	0%	0%	0%	0%	85%
	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	

Bob

$f_t^B = \mathcal{N}(\mu, \sigma^2)$

(climatological forecast)

Bob's reliability diagram



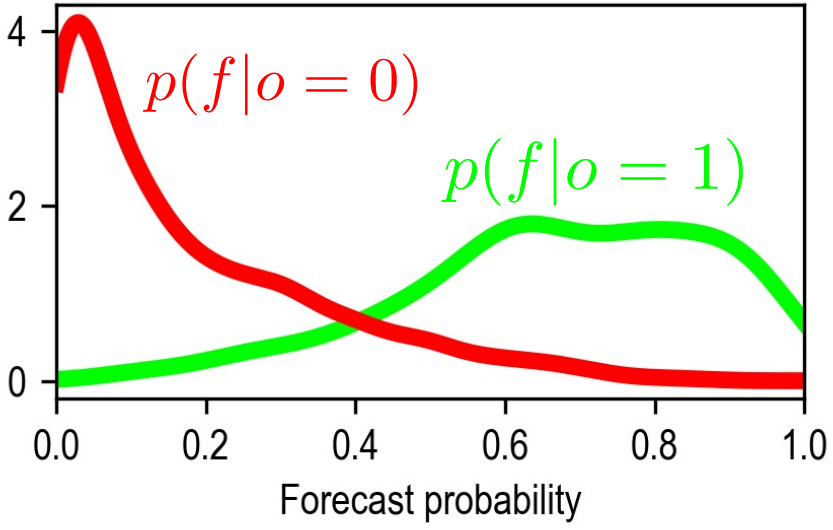
Bob's climatological forecast is perfectly reliable (but rather useless)

Discrimination: how verifying observations sort the forecast pdfs

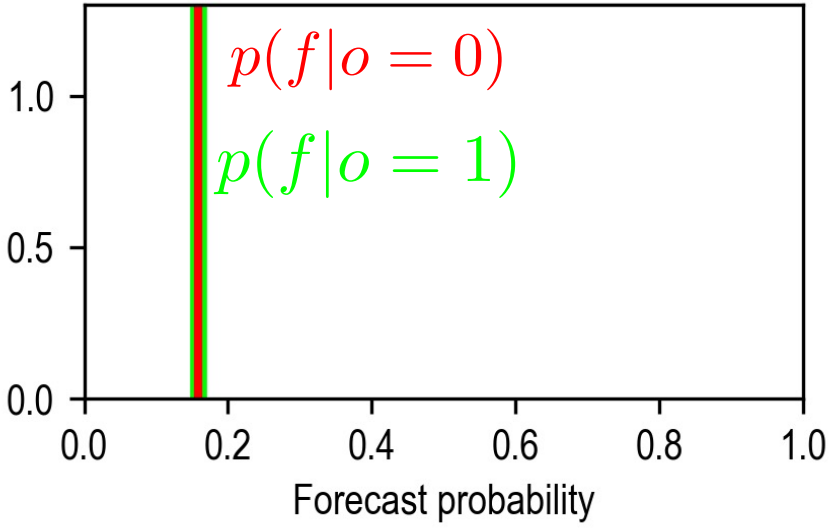
« Likelihood-base rate factorization »

$$p(f,o) = p(o|f)p(f) = p(f|o)p(o)$$

Alice: good discrimination



Bob: no discrimination

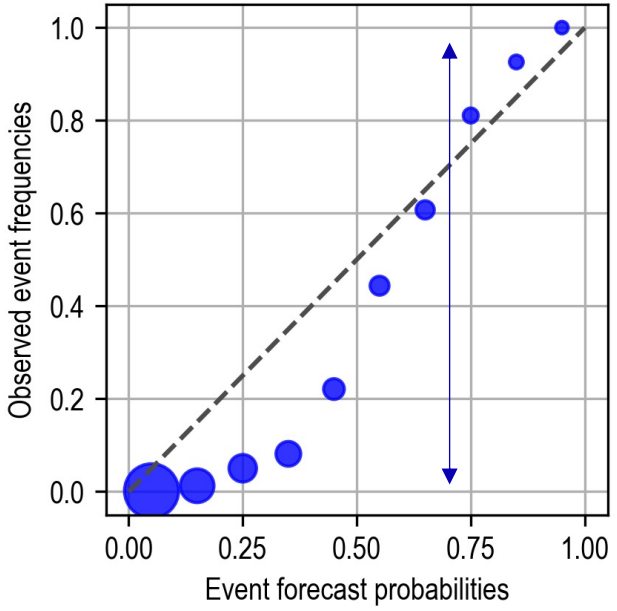


A forecast system is said to be exhibit **discrimination** if conditional forecast distributions differ depending on the observed outcome

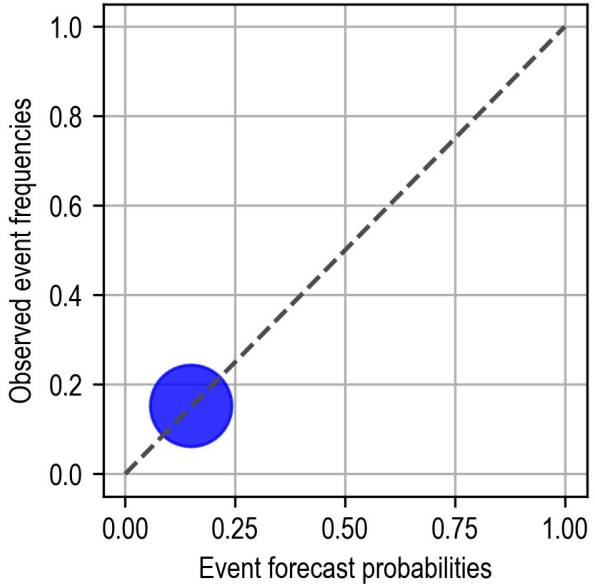
Resolution: how forecast probabilities sort observed frequencies

$$p(f,o) = p(o|f)p(f) = p(f|o)p(o)$$

Alice: good resolution



Bob: no resolution



A forecast system is said to be exhibit **resolution** if conditional observed frequencies differ depending on the forecast probabilities

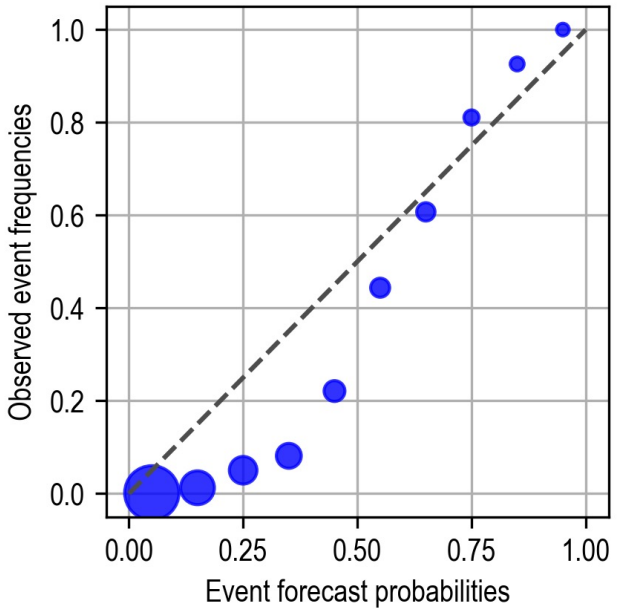
Resolution: how forecast probabilities sort observed frequencies

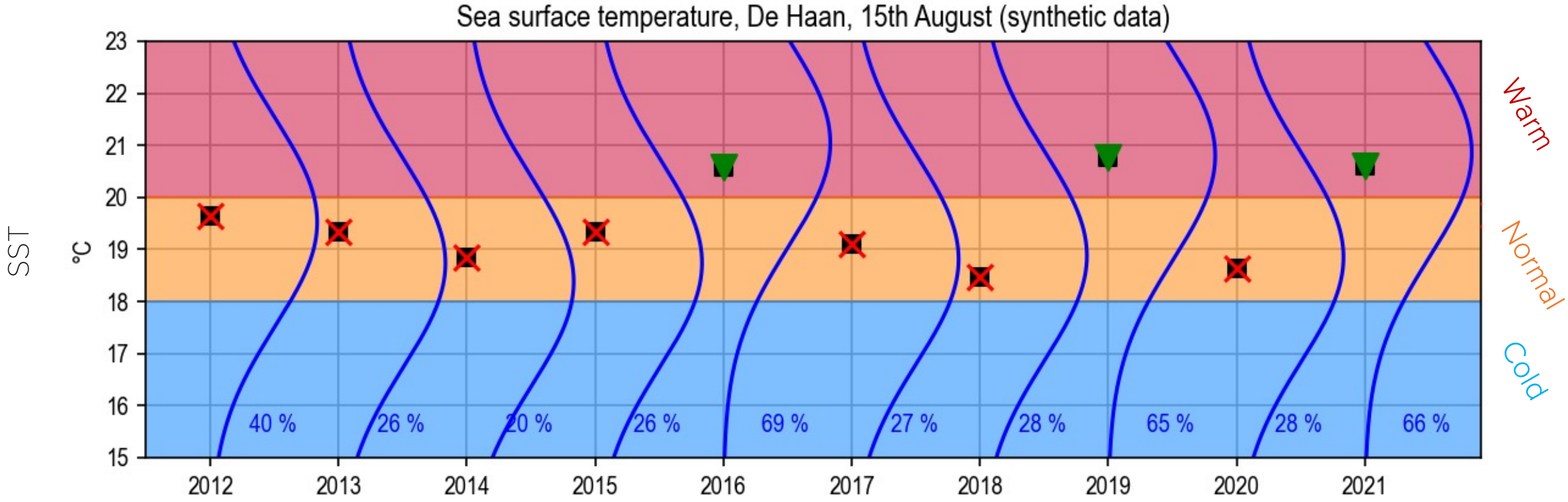
$$p(f,o) = p(o|f)p(f) = p(f|o)p(o)$$

Resolution

- is a property of the forecasts and the observations
- is not affected by the **statistical consistency** between forecast probabilities and observed frequencies. It is independent from reliability.
- pertains to the **differences** between the conditional distributions of observations for various forecast probabilities, while reliability compares them to the probabilities themselves
- **cannot** be improved by recalibration / post-processing of the forecasts

Alice's reliability diagram





$p(f^C)$

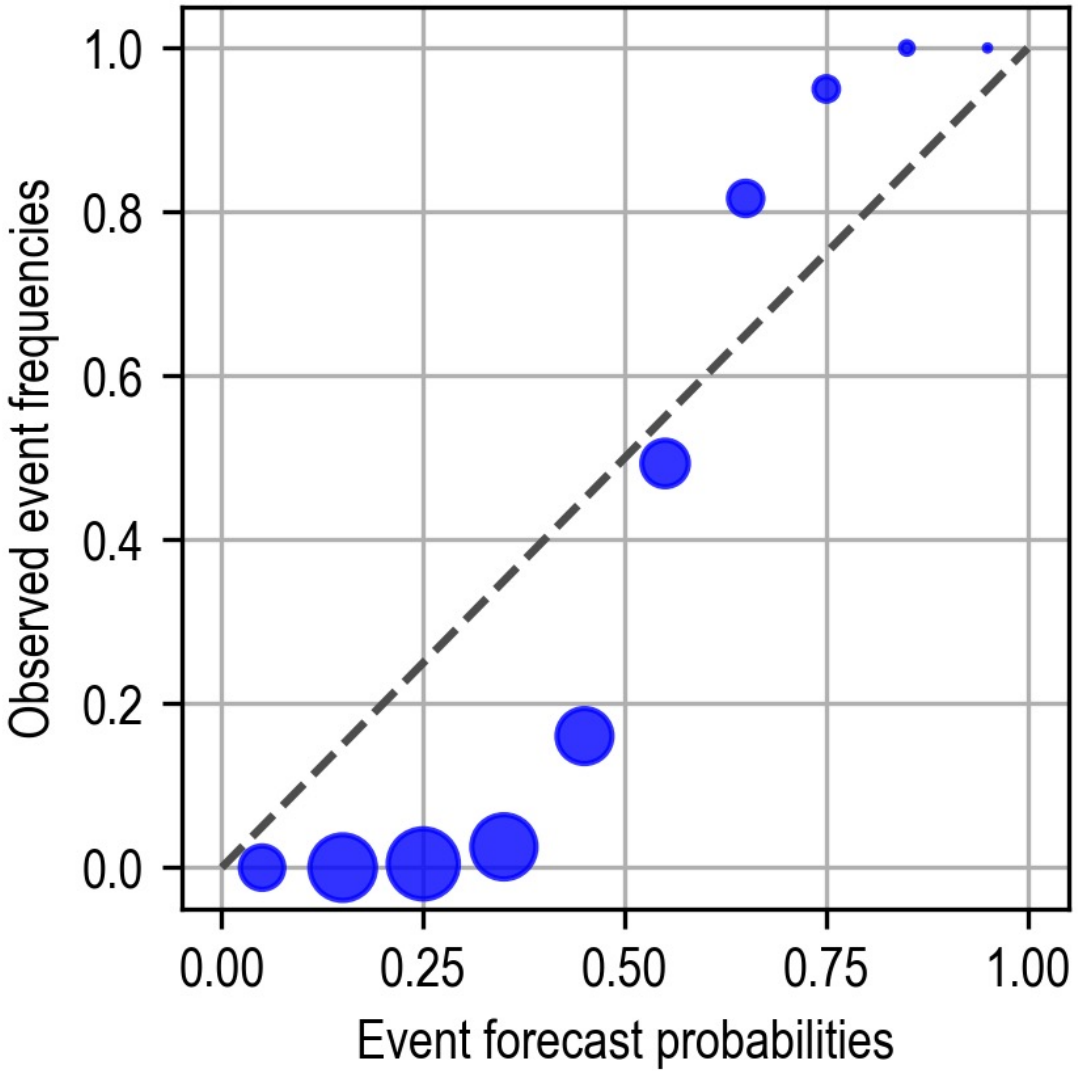
	0-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%	80-90%	90-100%	
$o = 1$	0%	0%	0%	0%	2%	5%	4%	3%	1%	0%	15%
$o = 0$	8%	19%	22%	18%	11%	5%	1%	0%	0%	0%	85%
	8%	19%	22%	18%	13%	10%	5%	3%	1%	0%	

Charles

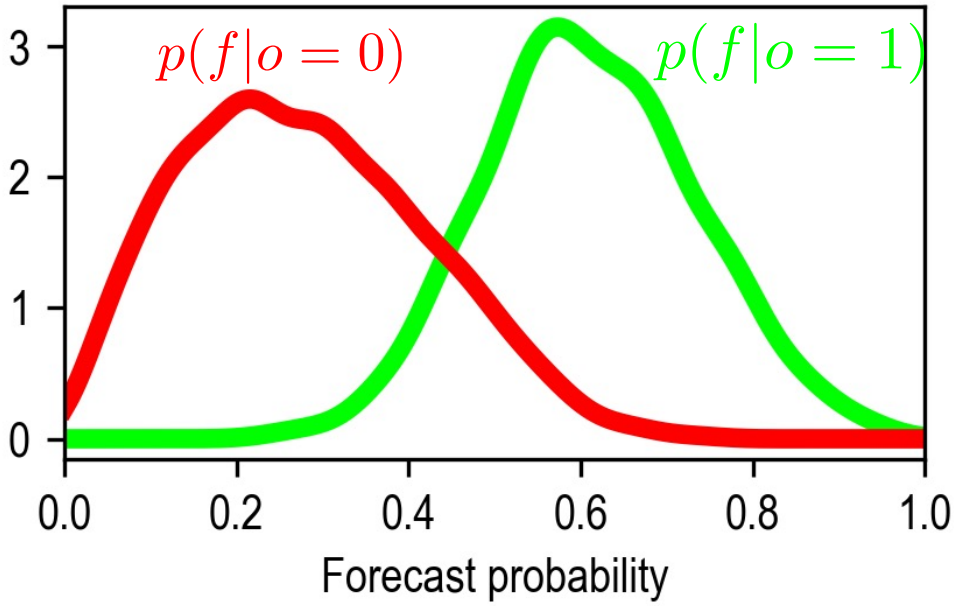
$$f_t^C = \mathcal{N}(\mu + \mathcal{N}(0, (0.5°\mathrm{C})^2, (2\sigma)^2)$$

(overdispersive/underconfident forecast)

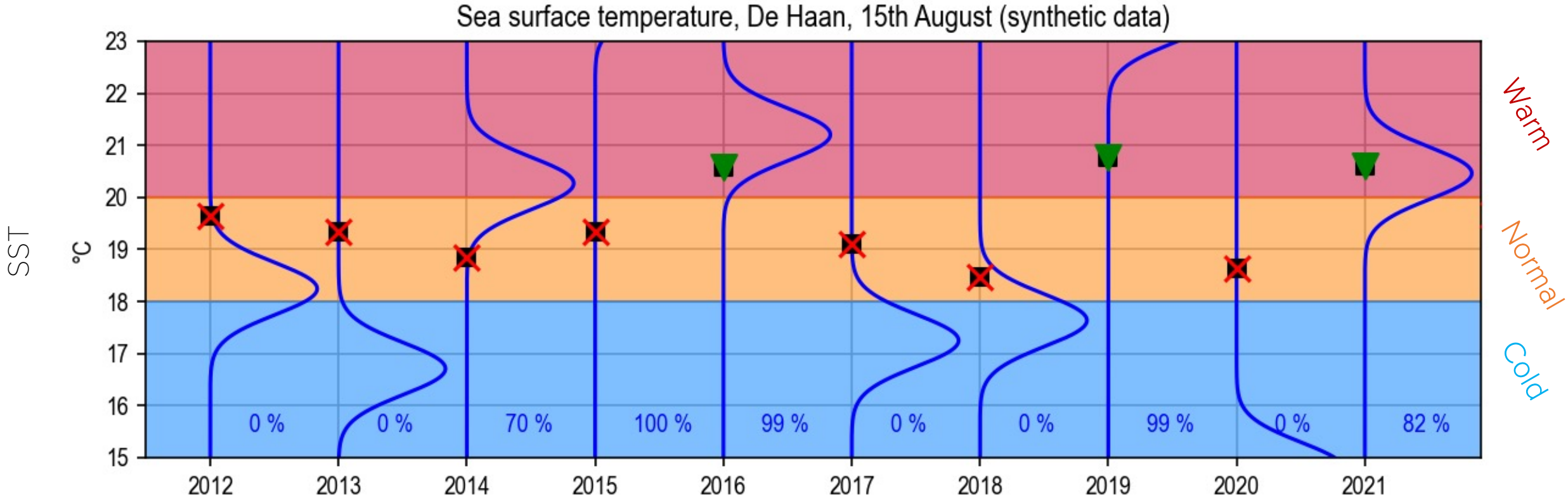
Charles's reliability diagram



Charles's discrimination



Poor reliability but
OK discrimination
and resolution



$p(f^D)$

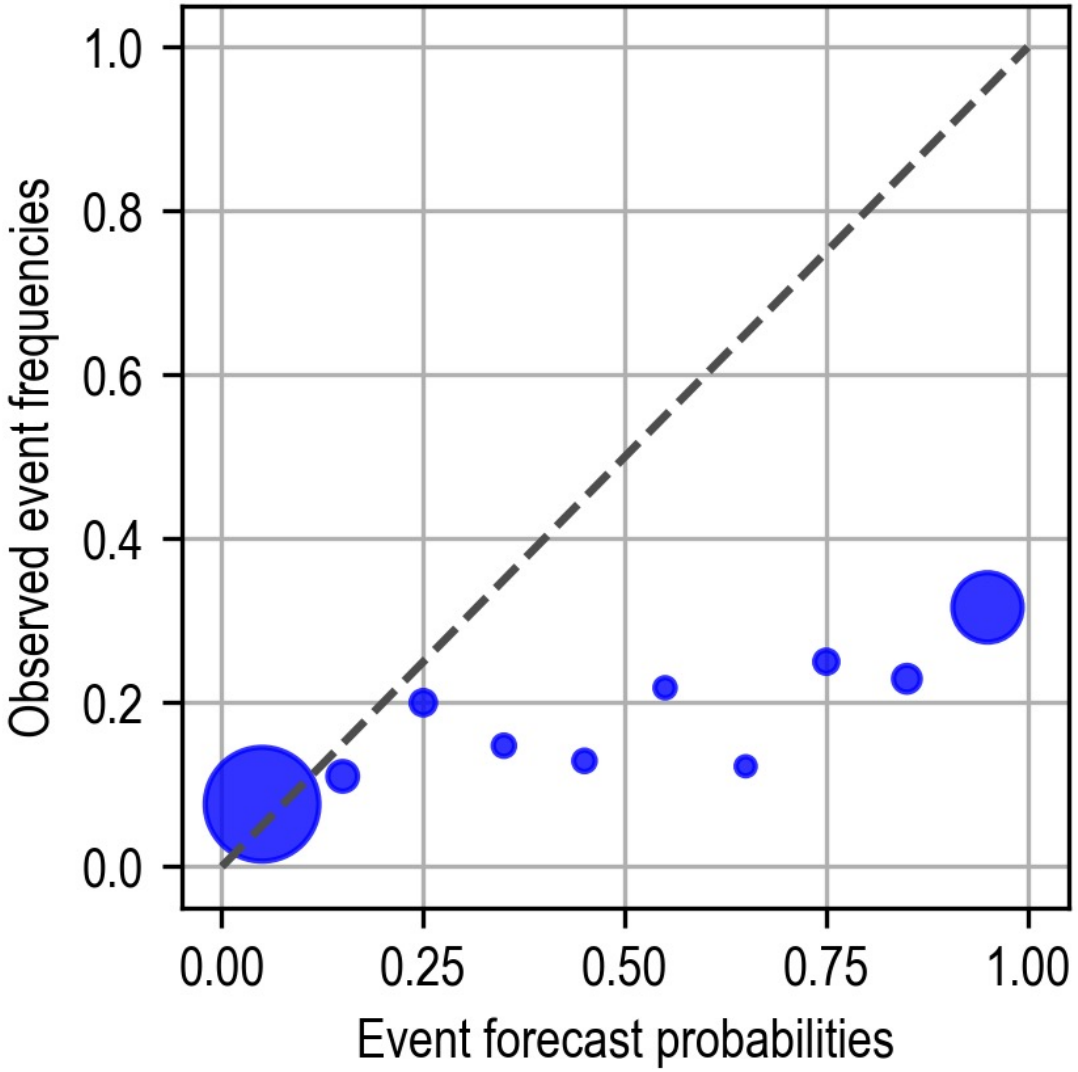
	0-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%	80-90%	90-100%	
$o = 1$	4%	0%	1%	0%	0%	0%	0%	1%	1%	7%	15%
$o = 0$	53%	4%	2%	2%	2%	1%	1%	2%	2%	15%	85%
	57%	4%	3%	2%	2%	1%	1%	3%	3%	22%	

Damien

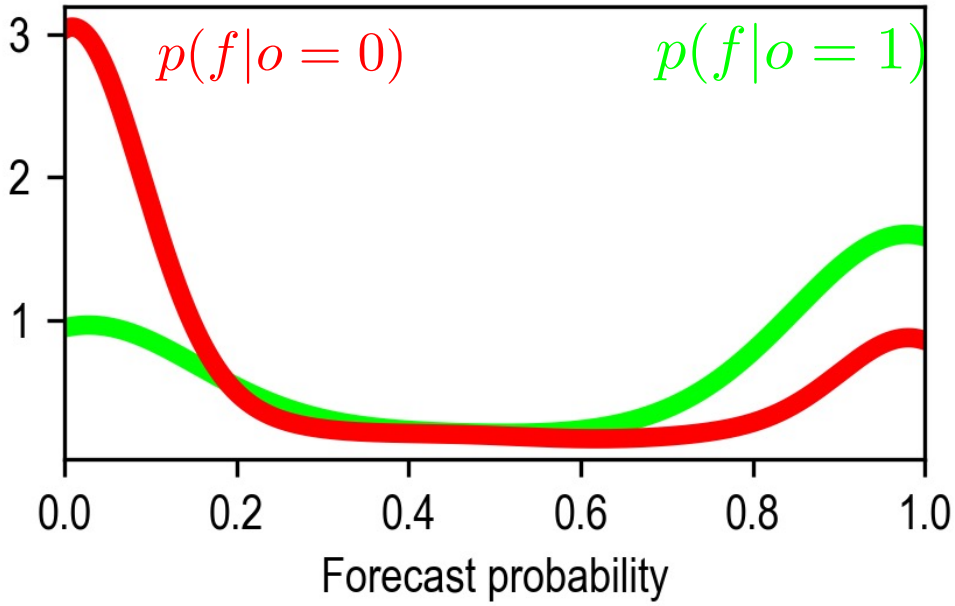
$$f_t^D = \mathcal{N}(\mu + \mathcal{N}(0, (2°\mathrm{C})^2, (0.5\sigma)^2)$$

(underdispersive/overconfident forecast)

Damien's reliability diagram



Damien's discrimination



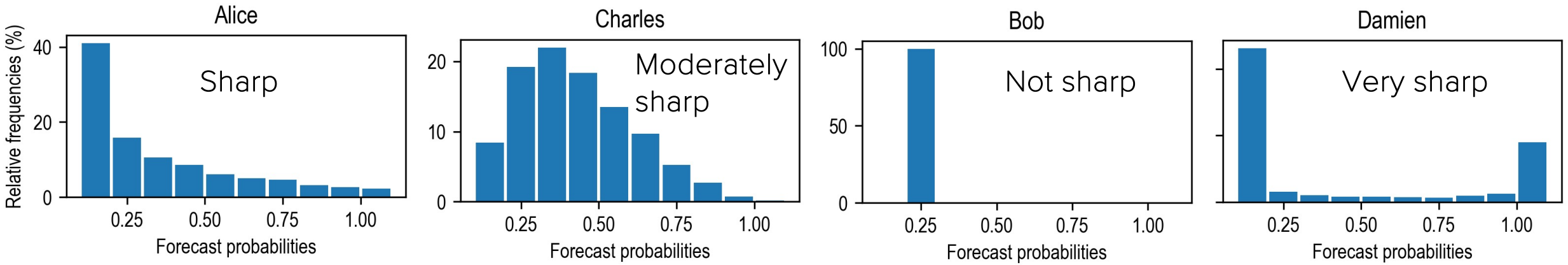
Poor reliability, OK discrimination, poor resolution

→ What is good in Damien's forecasts?

Sharpness (or refinement)

$$p(f,o) = p(o|f)p(f)$$

A forecast system is said to be **sharp** if its marginal (unconditional) forecast pdf differs significantly from the climatological pdf

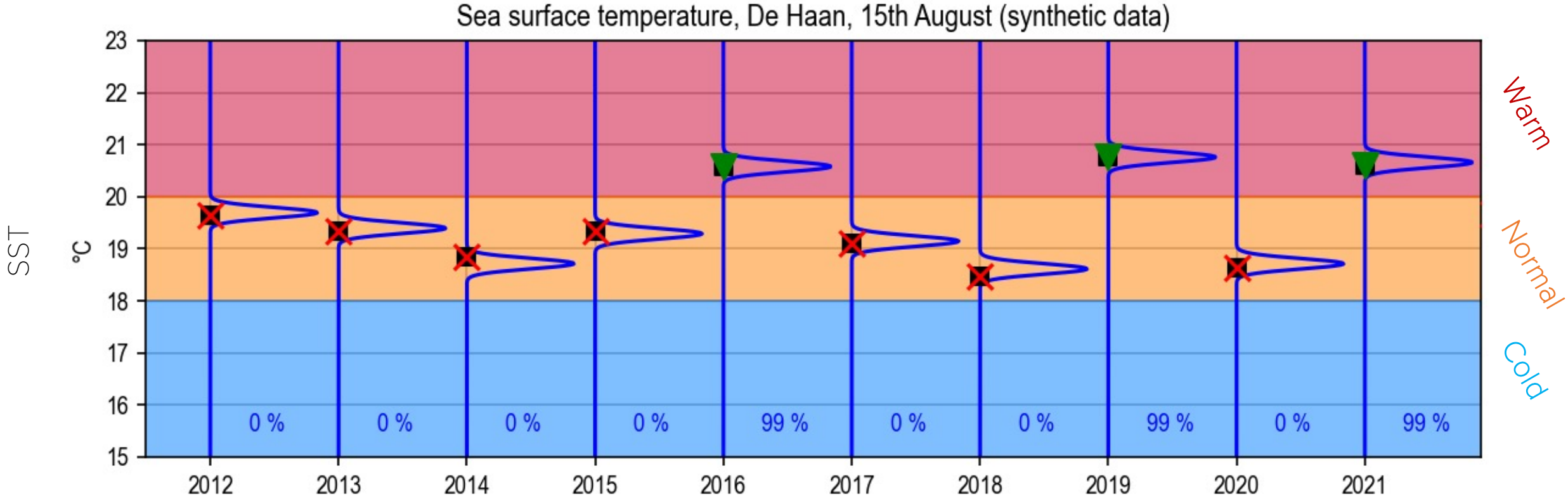


- Attribute of the forecast alone, regardless of the corresponding observations
- Anyone can produce sharp forecasts… but those are not necessarily reliable
- For a reliable forecast system, sharpness is identical to resolution

So what is the best forecast?

Reliability, sharpness, resolution and discrimination are called **‘attributes’** of a forecasting system.

The perfect forecast should ‘tick all the boxes’ (attributes)



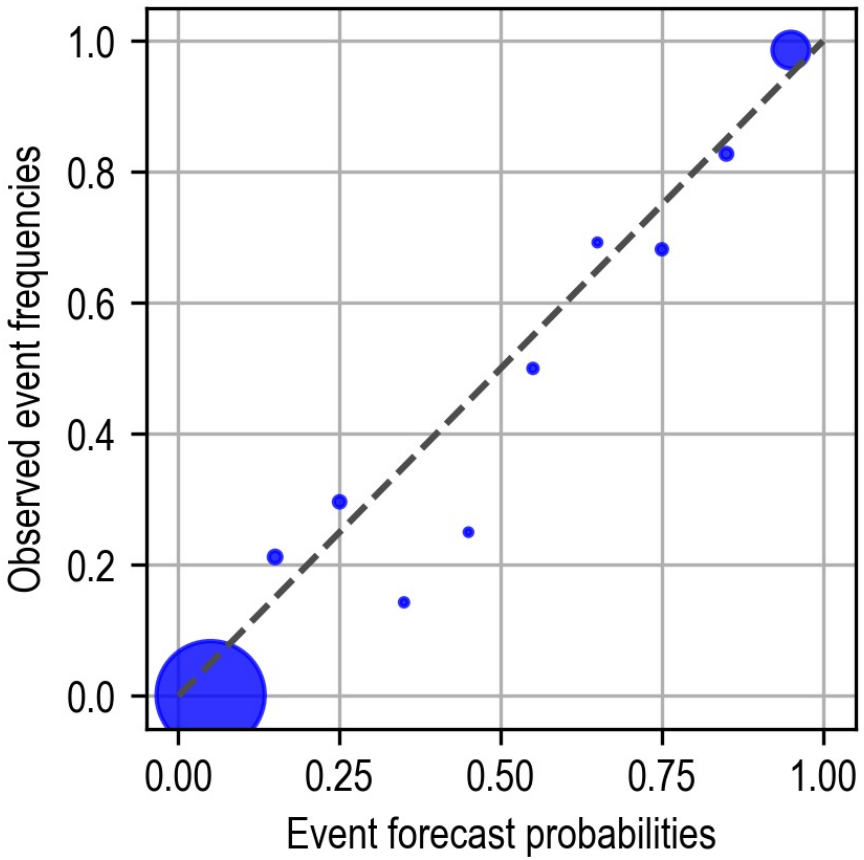
$p(f^D)$

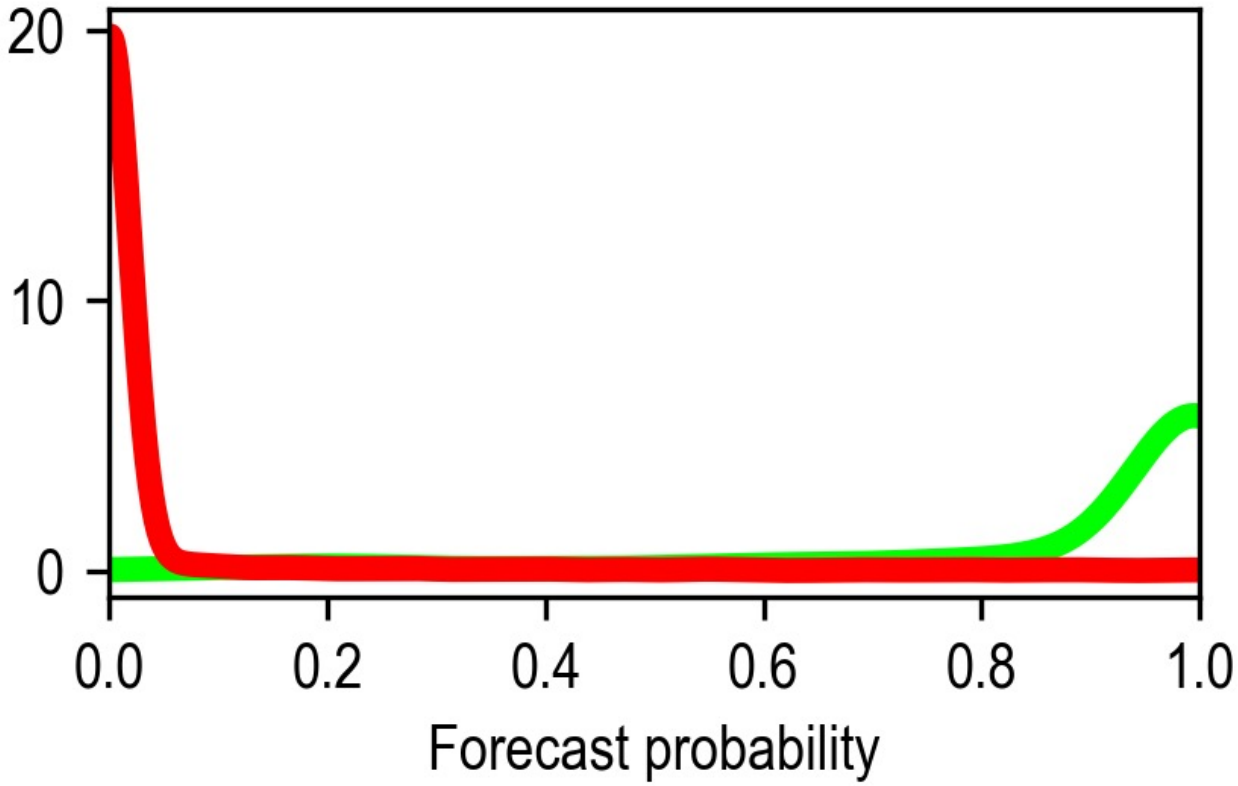
	0-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%	80-90%	90-100%	
$o = 1$	0%	0%	0%	0%	0%	0%	0%	0%	0%	9%	15%
$o = 0$	82%	1%	1%	0%	0%	0%	0%	0%	0%	0%	85%
	82%	1%	1%	0%	0%	1%	0%	1%	1%	9%	

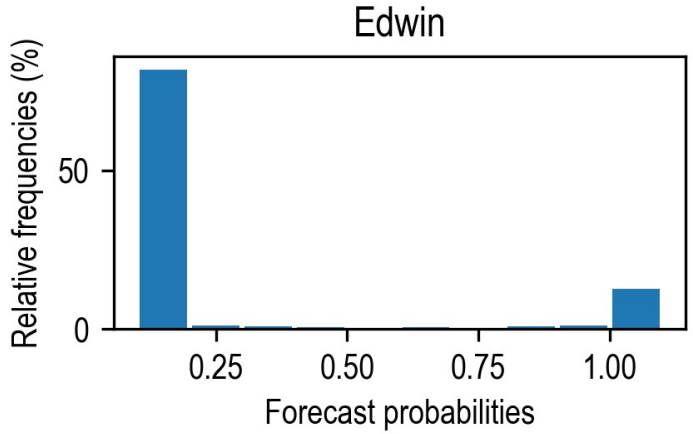
Edwin

$$f_t^E = \mathcal{N}(\mu + \mathcal{N}(0, (0.1°\mathrm{C})^2, (0.1\sigma)^2)$$

near-perfect forecast


Observed event frequencies
1.0
0.8
0.6
0.4
0.2
0.0
0.00
0.25
0.50
0.75
1.00
Event forecast probabilities


20
10
0
0.0
0.2
0.4
0.6
0.8
1.0
Forecast probability


Edwin
Relative frequencies (%)
50
0
0.25
0.50
0.75
1.00
Forecast probabilities

1. Historical perspective
2. Forecast verification defined
3. Will you go swimming this summer?
4. Scoring rules, scores and skill scores
5. Verification of full pdf forecasts
6. Visual representation of skill
7. Everything that was not said…

1. Historical perspective
2. Forecast verification defined
3. Will you go swimming this summer?
4. Scoring rules, scores and skill scores
5. Verification of full pdf forecasts
6. Visual representation of skill
7. Everything that was not said…

Scoring rules

A (probabilistic) scoring rule is a function that assigns a numerical value, or « score », to a given forecast pdf and a verifying observation:

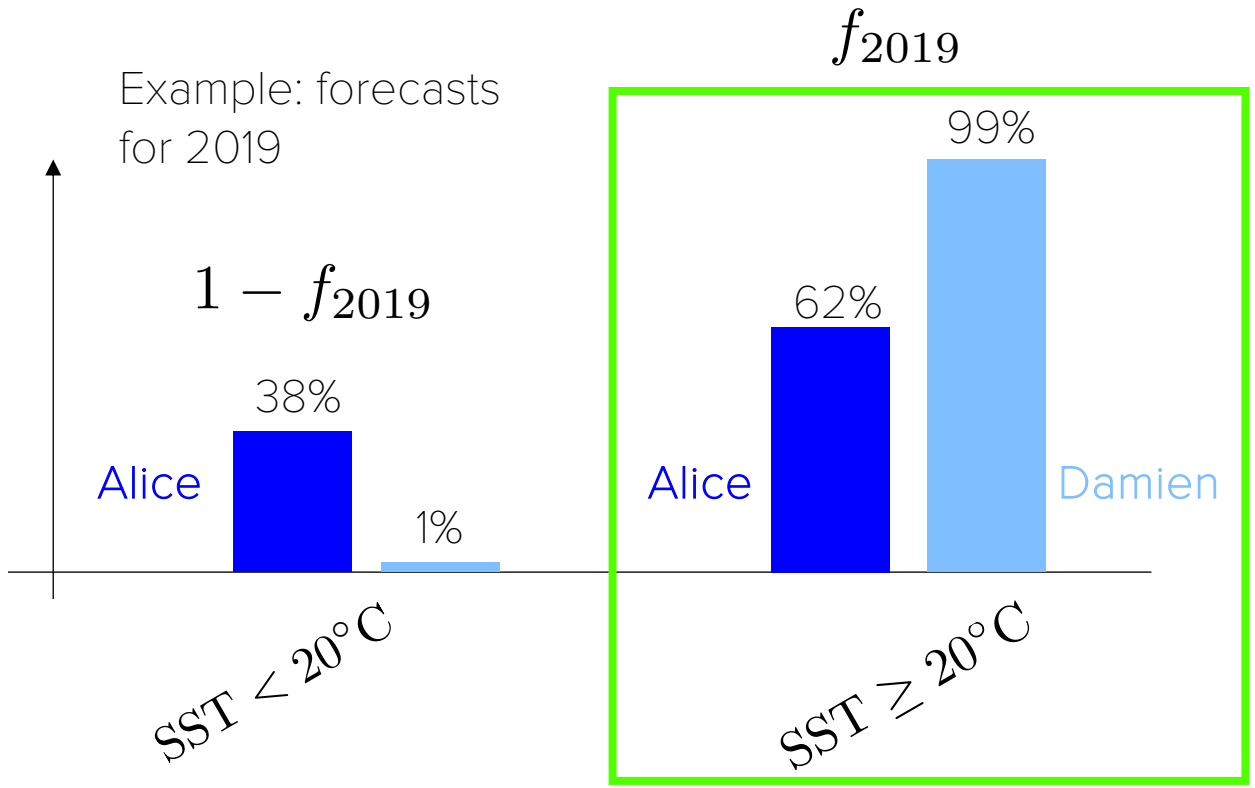
$$f, o \rightarrow s(f, o) \geq 0$$

By convention, lower = better

Gneiting, T., & Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437

The logarithmic (or « ignorance ») scoring rule (Good, 1952)

$$s(f_t, o_t) = \begin{cases} -\ln(f_t) & \text{if} \quad o_t = 1 \\ -\ln(1 - f_t) & \text{if} \quad o_t = 0 \end{cases}$$



$o_{2019} = 1$

Alice

$$-\ln(0.62) = 0.48$$

Damien

$$-\ln(0.99) = 0.01$$

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *14*(1), 107–114.

What is a « good » scoring rule?

$E : \mathrm{SST} \geq 20°\mathrm{C}$

$$E \rightarrow f \quad \rightarrow -\ln(f)$$
$$\neg E \rightarrow (1-f) \quad \rightarrow -\ln(1-f)$$

Expected gain: $\lambda = f(-\ln(f)) - (1-f)\ln(1-f)$

Suppose now the forecaster wants to 'play the system' by issuing a forecast probability that differs from his/her true belief (« hedging »)

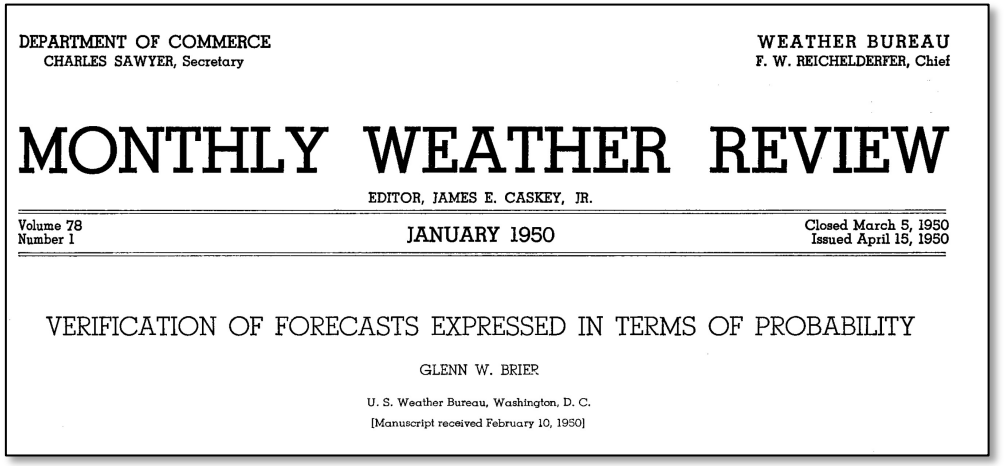
$$\lambda = f(-\ln(p)) - (1-f)\ln(1-p)$$

What is the p that maximizes his/her gains?

$$\frac{\partial \lambda}{\partial p} = 0 \Leftrightarrow p = f$$

Proper scoring rule

A scoring rule is said to be **proper** if the if the forecaster's expected score is minimized (i.e., is the best possible) when the forecaster issues a forecast pdf that is **equal to its best judgment**



DEPARTMENT OF COMMERCE
CHARLES SAWYER, Secretary

WEATHER BUREAU
F. W. REICHELDERFER, Chief

MONTHLY WEATHER REVIEW

EDITOR, JAMES E. CASKEY, JR.

Volume 78
Number 1

JANUARY 1950

Closed March 5, 1950
Issued April 15, 1950

VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY

GLENN W. BRIER

U. S. Weather Bureau, Washington, D. C.

[Manuscript received February 10, 1950]

« It is the purpose of this paper to discuss one situation where it appears to be possible to **devise a verification scheme** that **cannot influence the forecaster in any undesirable way** »

Proper scoring rules prevent forecasters from « hedging » or « playing the system »

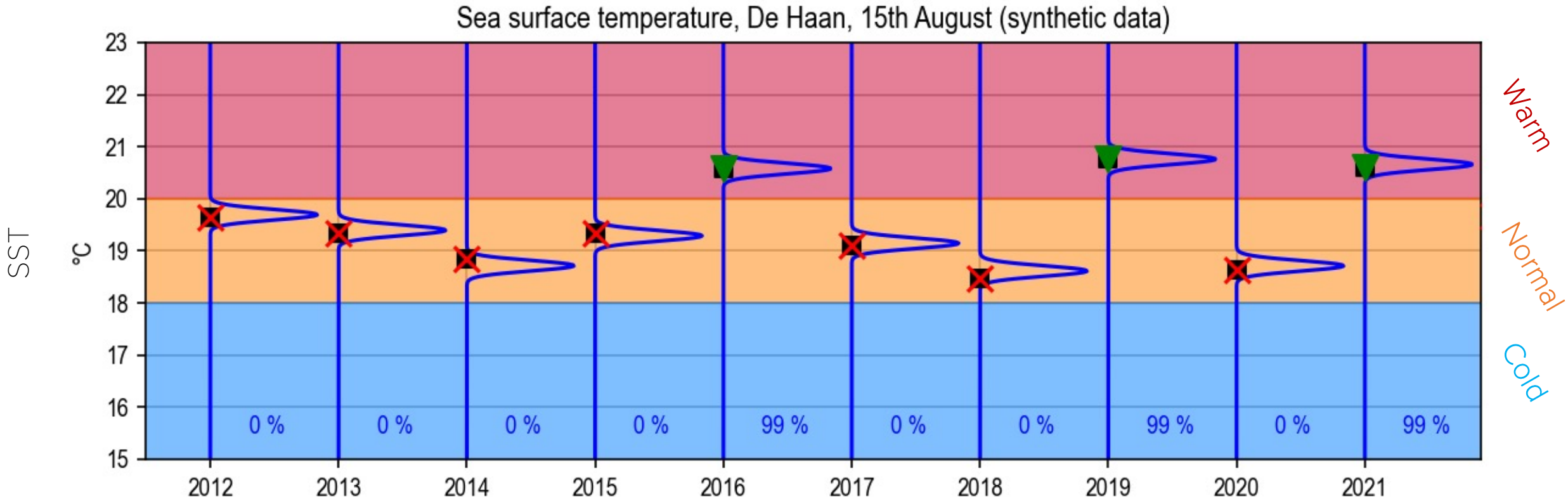
Examples of scoring rules: logarithmic, Brier, CRPS, spherical (see later)

The Brier Score

$$\mathrm{BS} = \frac{1}{T}\sum_{t=1}^{T}(f_t - o_t)^2$$

The BS hits hard « middle-of-the road » forecasters

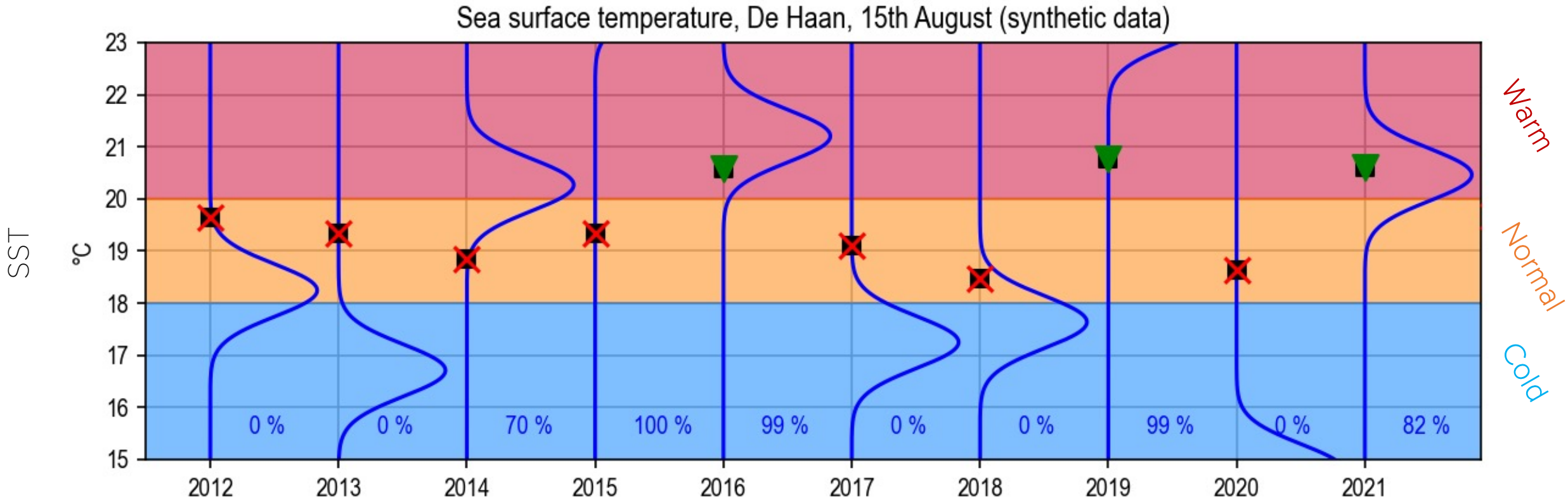
… but hits even harder confident forecasters that turn out to be on the wrong side of the event!



Edwin

BS = 0.01

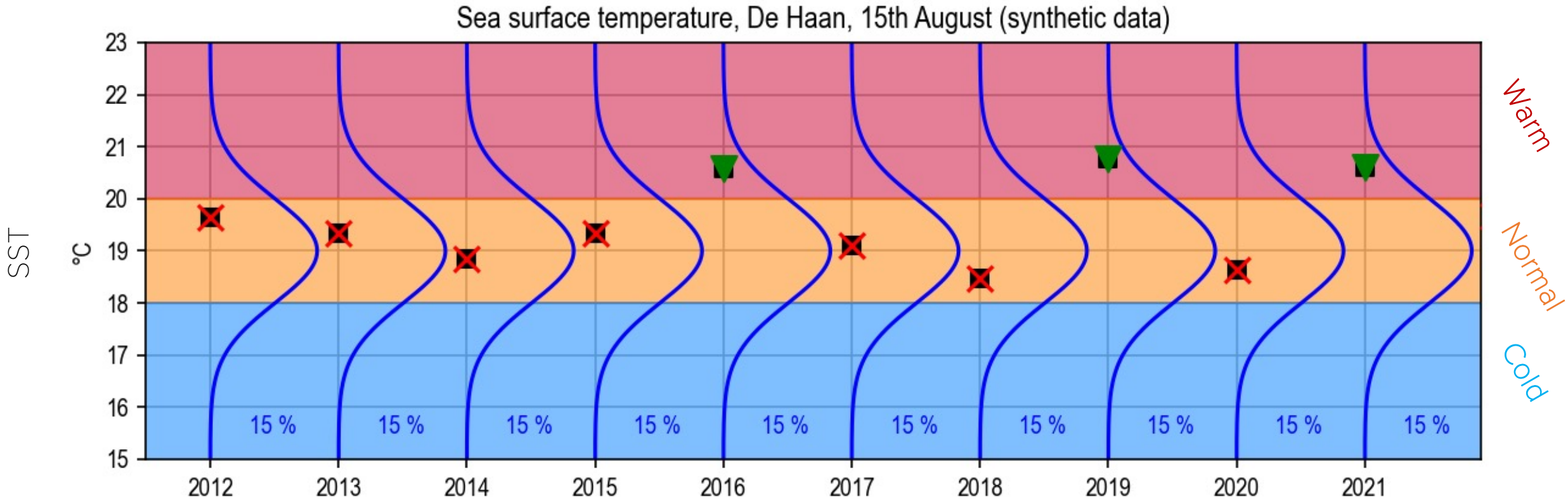
$$\mathrm{BS} = \frac{1}{T}\sum_{t=1}^{T}(f_t - o_t)^2$$



Edwin
BS = 0.01

Damien
BS = 0.25

$$\mathrm{BS} = \frac{1}{T}\sum_{t=1}^{T}(f_t - o_t)^2$$

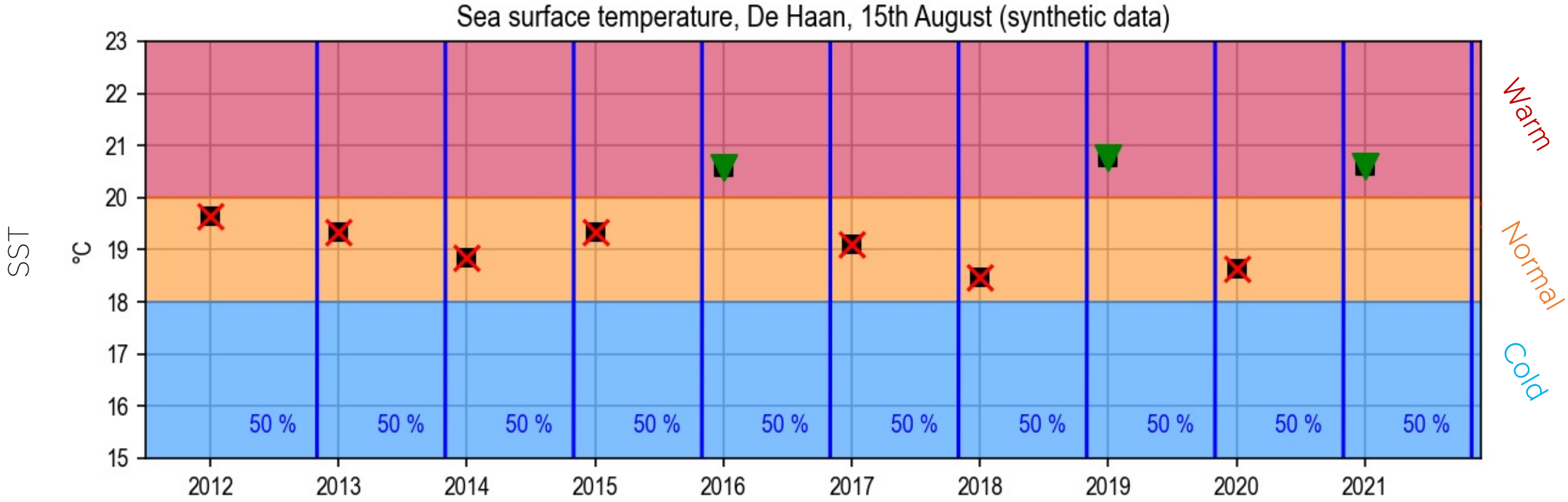


Edwin
BS = 0.01

Damien
BS = 0.25

Bob
BS = 0.13

$$\mathrm{BS} = \frac{1}{T}\sum_{t=1}^{T}(f_t - o_t)^2$$



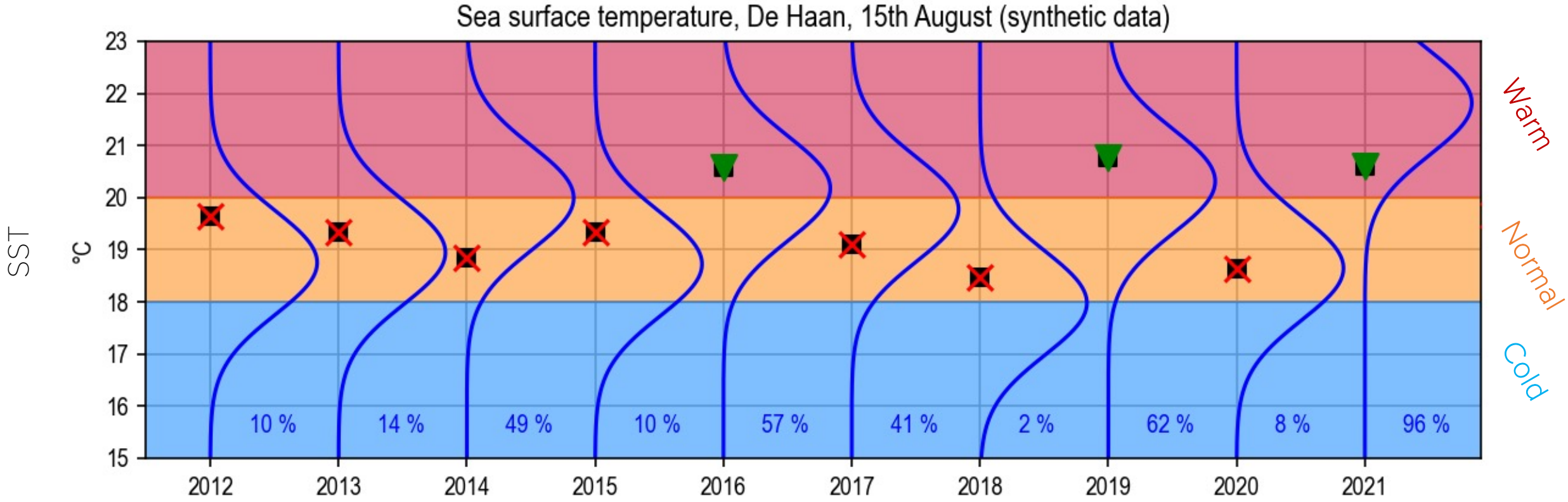
Edwin
BS = 0.01

Damien
BS = 0.25

Bob
BS = 0.13

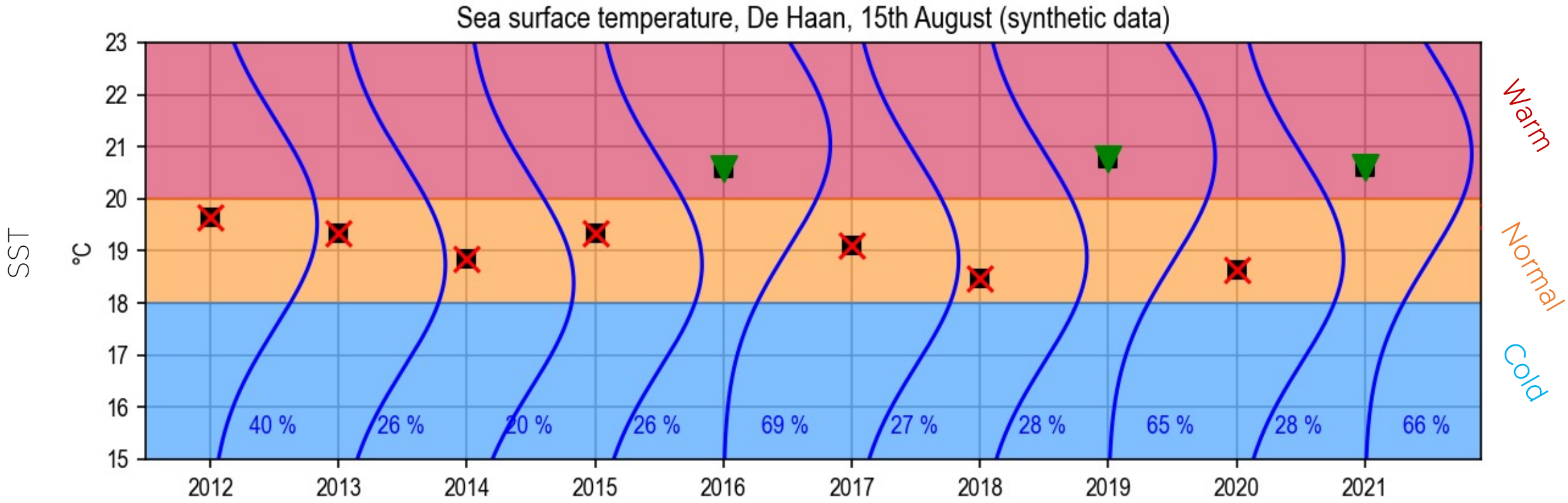
50-50
BS = 0.25

$$\mathrm{BS} = \frac{1}{T}\sum_{t=1}^{T}(f_t - o_t)^2$$



Edwin	Damien	Bob	50-50	Alice
BS = 0.01	BS = 0.25	BS = 0.13	BS = 0.25	BS = 0.13

$$\mathrm{BS} = \frac{1}{T}\sum_{t=1}^{T}(f_t - o_t)^2$$



Edwin	Damien	Bob	50-50	Alice	Charles
BS = 0.01	BS = 0.25	BS = 0.13	BS = 0.25	BS = 0.13	BS = 0.11

$$\mathrm{BS} = \frac{1}{T}\sum_{t=1}^{T}(f_t - o_t)^2$$

The Brier Scores of trivial forecasts

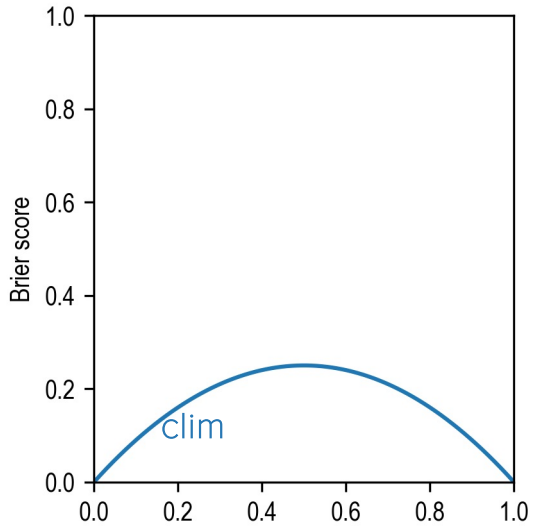
- Let the baseline rate (or climatological frequency of occurrence) of the event be $\overline{o} = \frac{1}{T}\sum_{t=1}^{T} o_t$

Cheap forecast strategies:

1) Forecast the baseline rate
(« Climatological forecast »)

$$\rightarrow BS = \overline{o}(1-\overline{o})$$

Very rewarding strategy for low-probability events (cf. Finley)



The Brier Scores of trivial forecasts

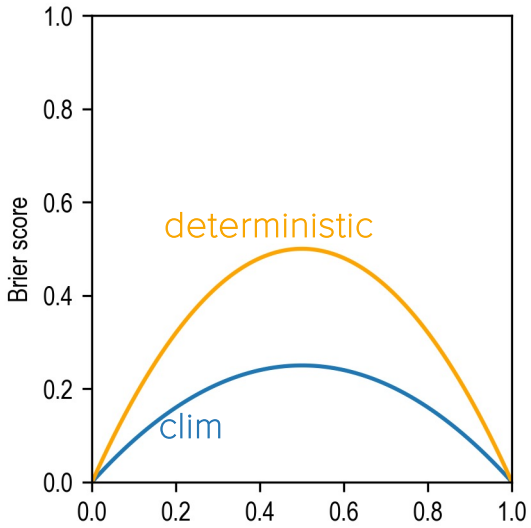
- Let the baseline rate (or climatological frequency of occurrence) of the event be $\overline{o} = \frac{1}{T}\sum_{t=1}^{T} o_t$

Cheap forecast strategies:

2) Deterministic:
Forecast 100% with probability $\overline{o}$, 0% else

$$\rightarrow BS = 2\overline{o}(1-\overline{o})$$

Penalizes systematic over confidence

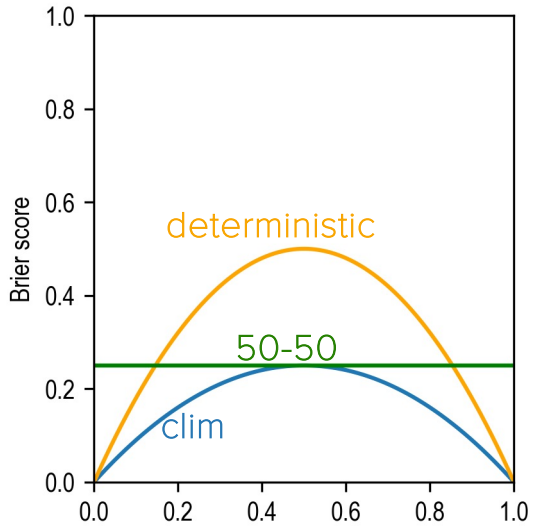


The Brier Scores of trivial forecasts

- Let the baseline rate (or climatological frequency of occurrence) of the event be $\overline{o} = \frac{1}{T}\sum_{t=1}^{T} o_t$

Cheap forecast strategies:

3) 50-50 forecast $\rightarrow BS = 0.25$



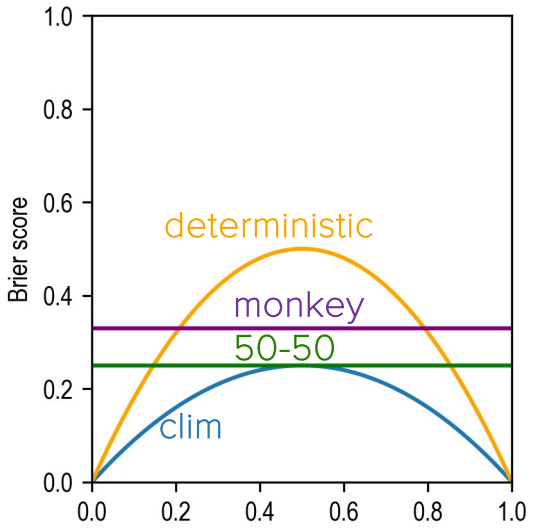
The Brier Scores of trivial forecasts

- Let the baseline rate (or climatological frequency of occurrence) of the event be $\overline{o} = \frac{1}{T}\sum_{t=1}^{T} o_t$

Cheap forecast strategies:

4) 'Monkey forecast'
Random probability in [0, 1]

$$\rightarrow BS = 0.33$$



Brier score decomposition

$$BS = \frac{1}{T}\sum_{i=1}^{I} n^i (f_t^i - \overline{o}^i)^2 - \frac{1}{T}\sum_{i=1}^{I} n^i (\overline{o}^i - \overline{o})^2 + \underbrace{\overline{o}(1-\overline{o})}$$

Whether forecast probabilities are matched by observed frequencies

Whether the conditional observed frequencies differ from the unconditional base rate

BS of climatological forecast!

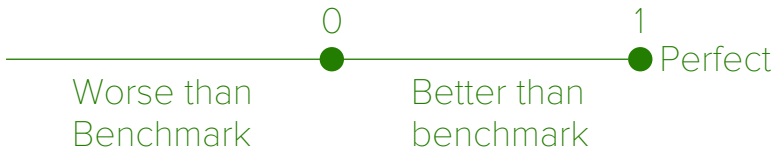
$$= REL - RES + UNC$$

Skill scores

« Brier Skill Score »:

$$BSS = \frac{BS - BS_{\text{ref}}}{0 - BS_{\text{ref}}} = 1 - \frac{BS}{BS_{\text{ref}}}$$

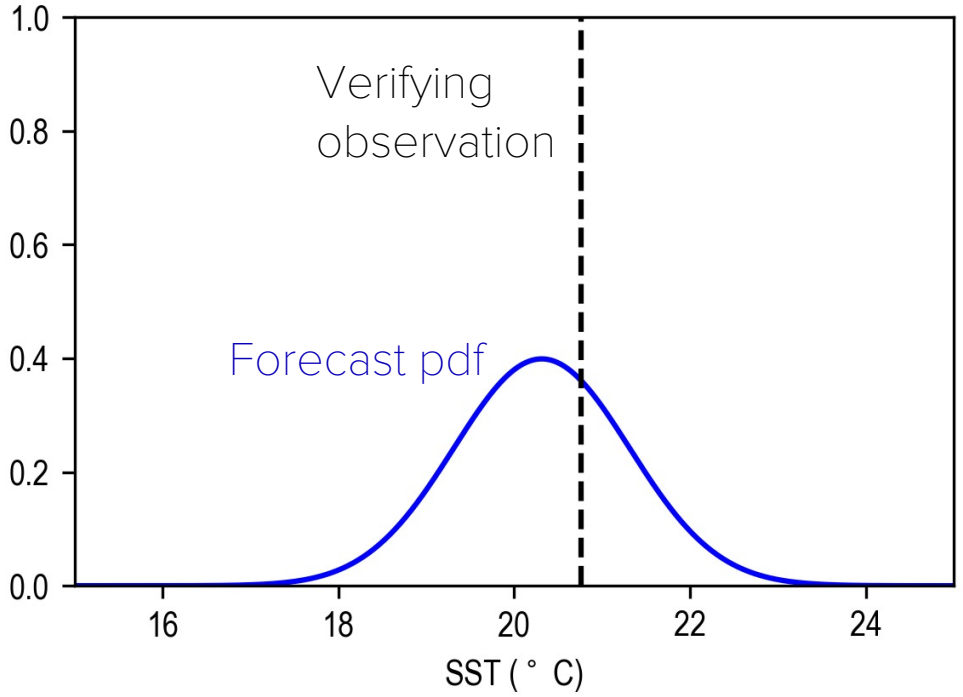
Re-scales the BS according to a reference forecast (e.g., climatological)



1. Historical perspective
2. Forecast verification defined
3. Will you go swimming this summer?
4. Scoring rules, scores and skill scores
5. Verification of full pdf forecasts
6. Visual representation of skill
7. Everything that was not said…

1. Historical perspective
2. Forecast verification defined
3. Will you go swimming this summer?
4. Scoring rules, scores and skill scores
5. Verification of full pdf forecasts
6. Visual representation of skill
7. Everything that was not said…

The continuous rank probability score

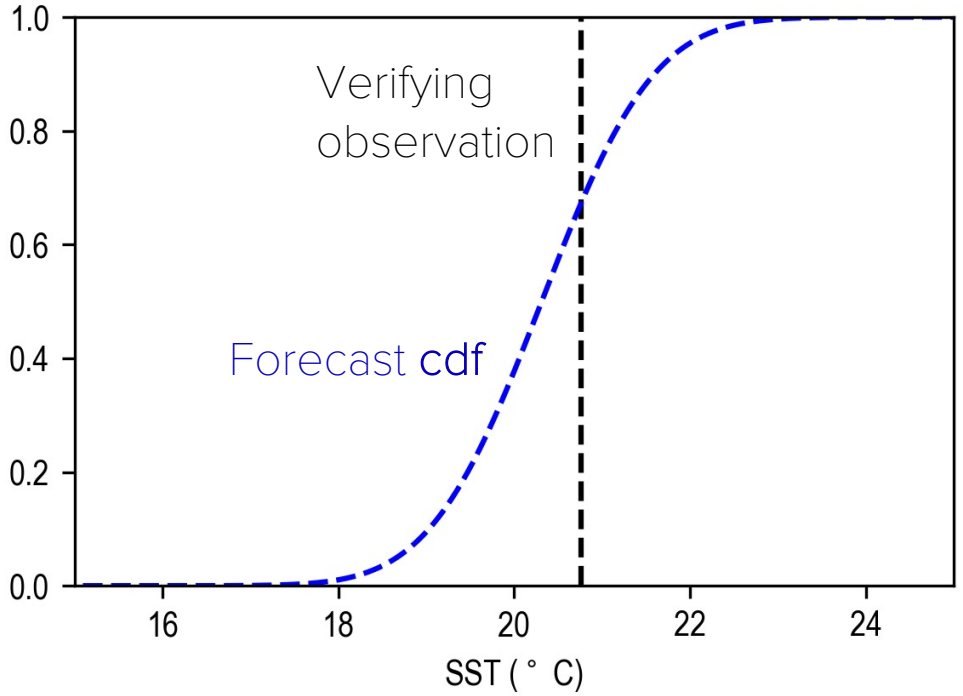


The continuous rank probability score

$$CRPS = \int_{-\infty}^{\infty} (F_f(x) - F_o(x))^2 \, \mathrm{d}x$$

cdf of forecasts

cdf of the verifying observations, i.e., step function at the observed value

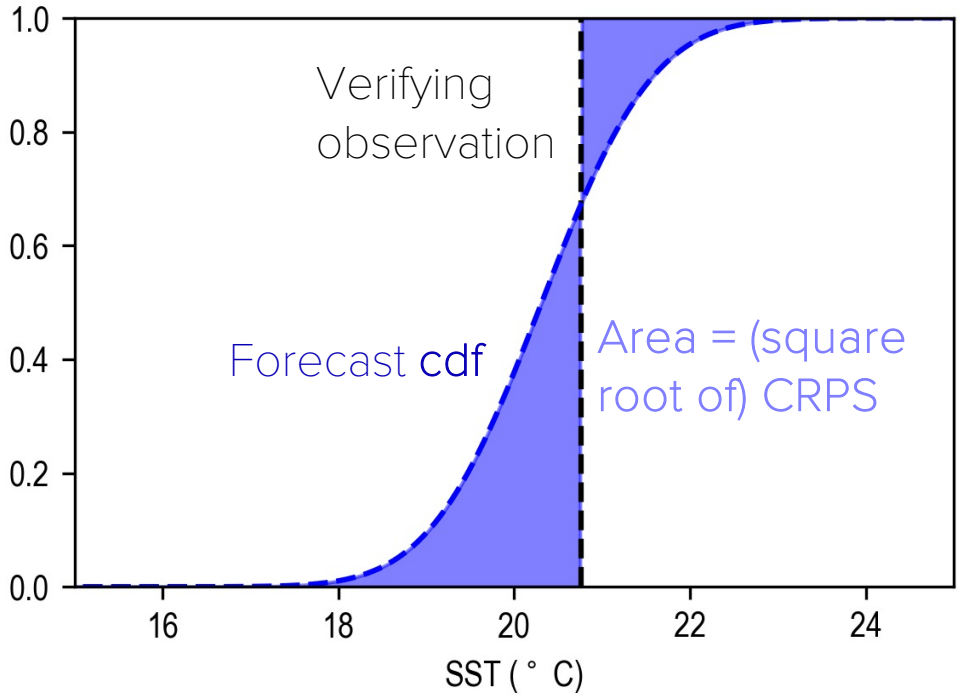


The continuous rank probability score

$$CRPS = \int_{-\infty}^{\infty} (F_f(x) - F_o(x))^2 \, \mathrm{d}x$$

cdf of forecasts

cdf of the verifying observations, i.e., step function at the observed value



The CRPS is a proper score and can be partitioned just like the Brier Score (Hersbach, 2000)

Hersbach, H. (2000). Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. Weather and Forecasting, 15(5), 559–570. https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2

1. Historical perspective
2. Forecast verification defined
3. Will you go swimming this summer?
4. Scoring rules, scores and skill scores
5. Verification of full pdf forecasts
6. Visual representation of skill
7. Everything that was not said…

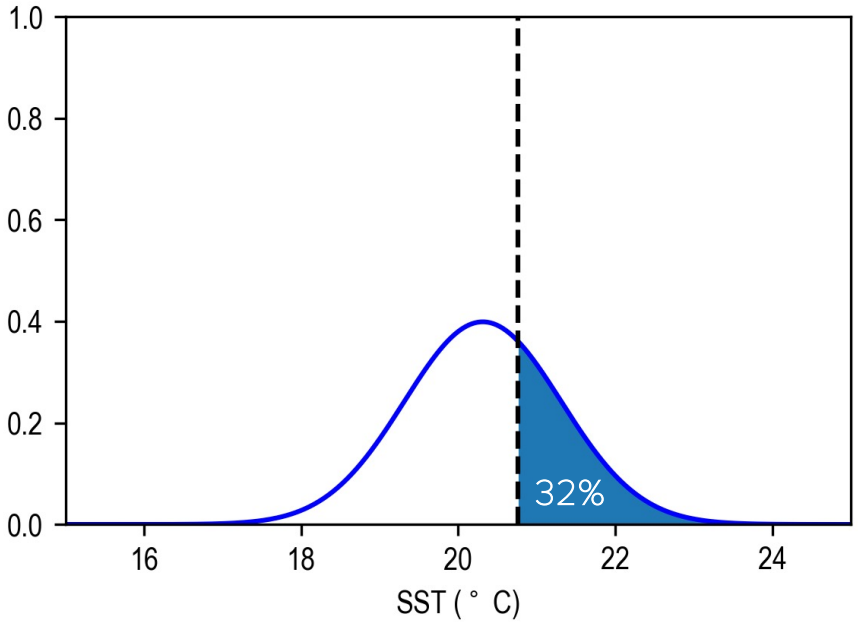
1. Historical perspective
2. Forecast verification defined
3. Will you go swimming this summer?
4. Scoring rules, scores and skill scores
5. Verification of full pdf forecasts
6. Visual representation of skill
7. Everything that was not said…

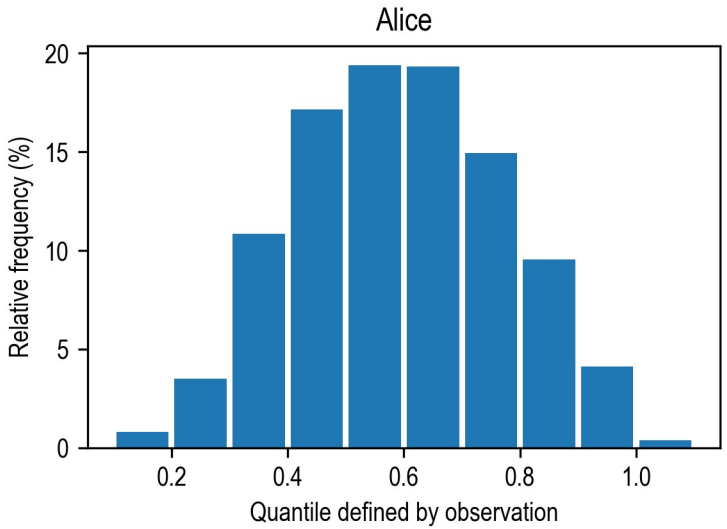
Rank histogram / Talagrand diagram

1. Locate the verifying observation in the forecast pdf in terms of quantile / rank

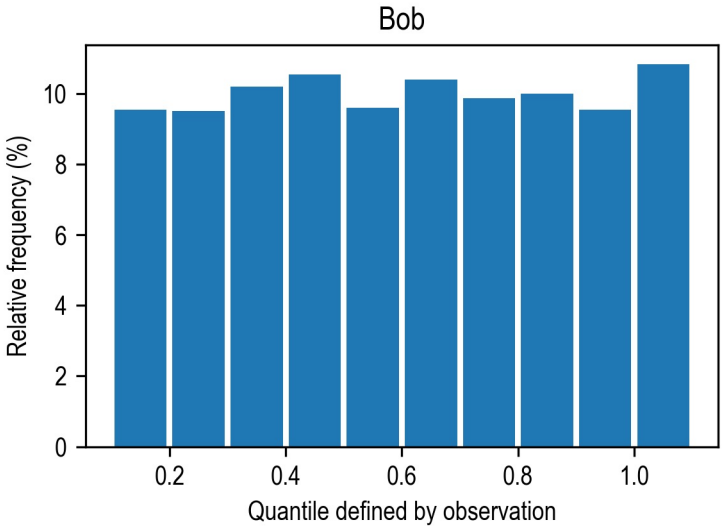
2. Construct the histogram over all forecasts of the obtained quantiles /ranks

If the forecast pdf is the same as the observed pdf, then all quantiles / ranks appear with the same frequency

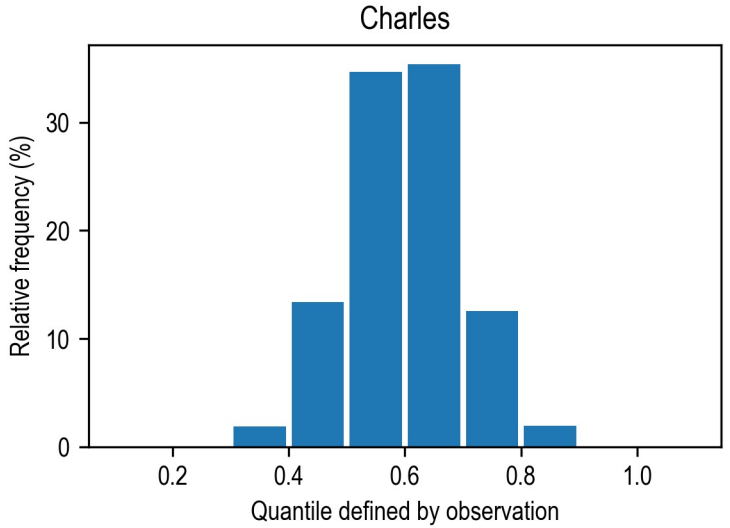




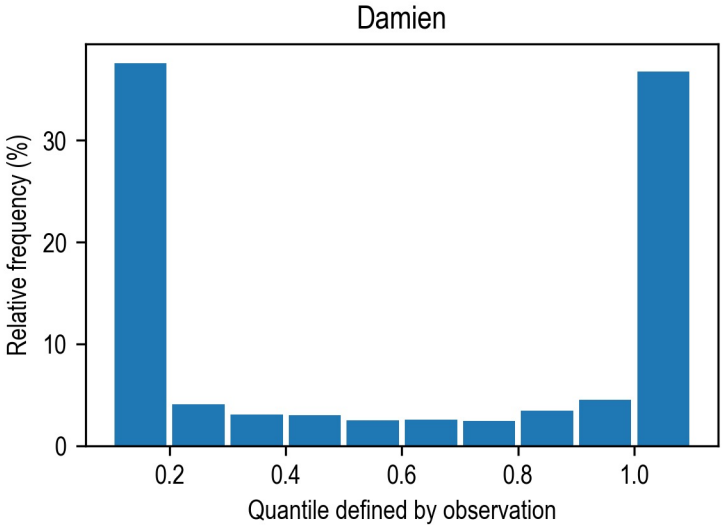
Under confident



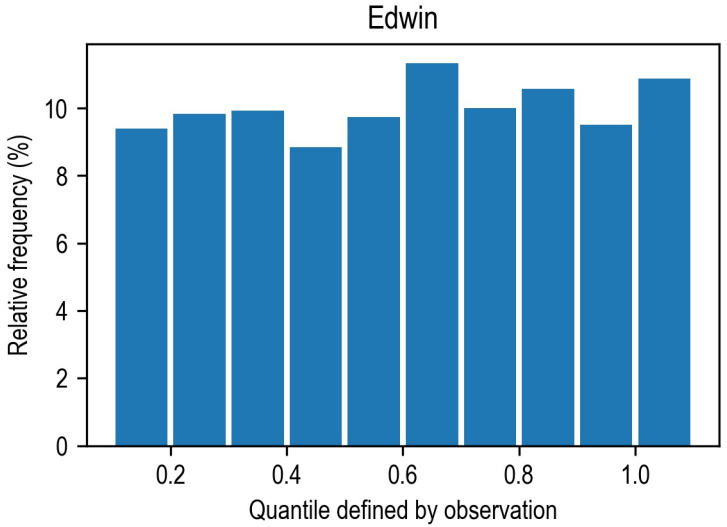
Perfect



Under-confident / overdispersive



Overconfident



Perfect

1. Historical perspective
2. Forecast verification defined
3. Will you go swimming this summer?
4. Scoring rules, scores and skill scores
5. Verification of full pdf forecasts
6. Visual representation of skill
7. Everything that was not said…

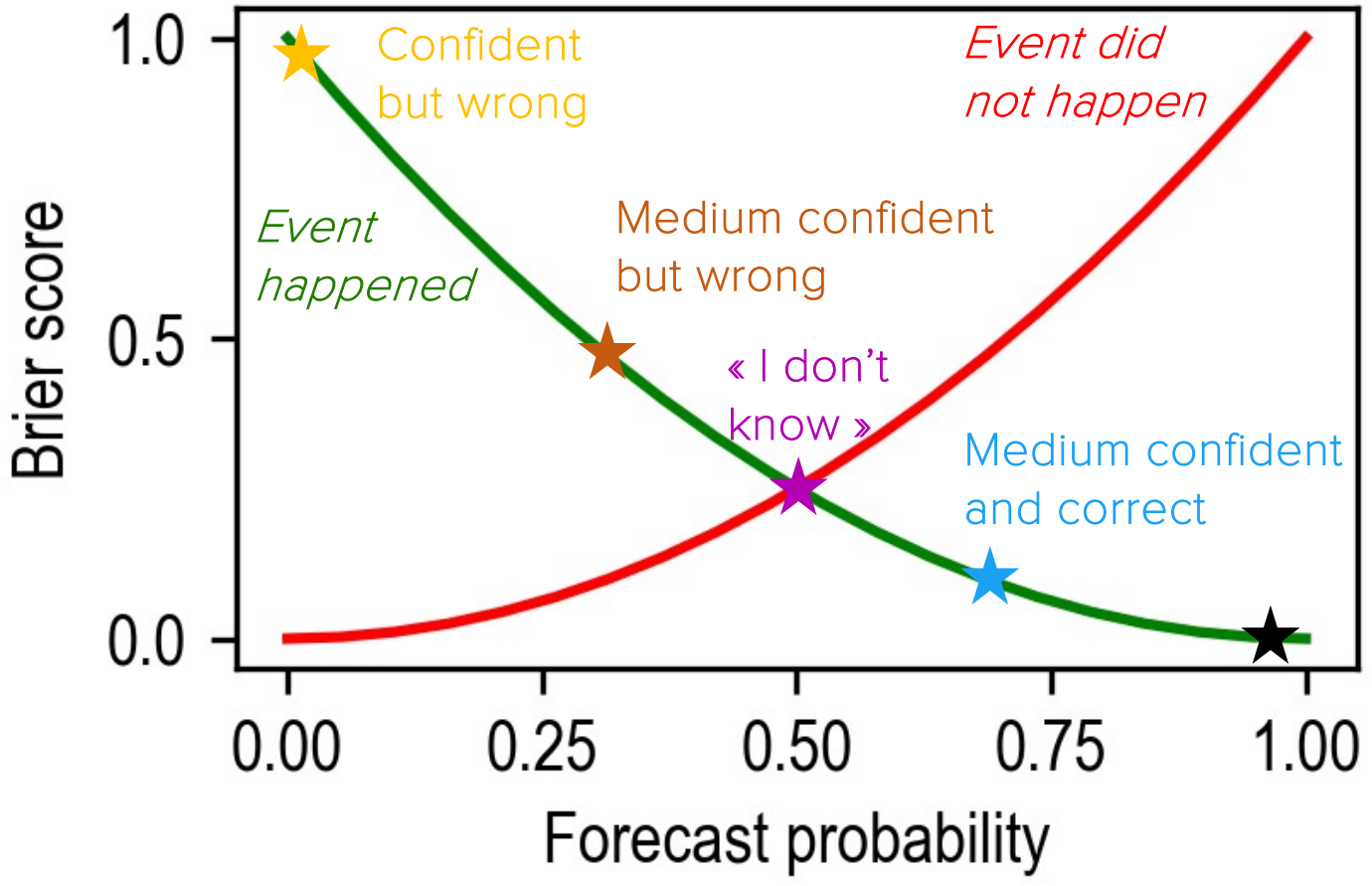
1. Historical perspective
2. Forecast verification defined
3. Will you go swimming this summer?
4. Scoring rules, scores and skill scores
5. Verification of full pdf forecasts
6. Visual representation of skill
7. Everything that was not said…

- Exact forecast pdfs are rarely available. Equiprobable realizations (ensemble members) are used instead → sampling issues
- Robust statistics require long-term series (large T). This is not the case with climate. Still, the same scores can be used for skill intercomparison
- The scores implicitly assume some stationarity in the verification data → obviously not the case in a climate context
- Recent research aims at including observational uncertainty into these forecast verification metrics (e.g., Naveau and Bessac, 2018; Ferro, 2017)
- Forecast verification theory trains us to think probabilistically

Naveau, P., & Bessac, J. (2018). Forecast evaluation with imperfect observations and imperfect models. ArXiv:1806.03745 [Stat]. http://arxiv.org/abs/1806.03745

Ferro, C. A. T. (2017). Measuring forecast performance in the presence of observation error: Measuring Forecast Performance in the Presence of Observation Error. *Quarterly Journal of the Royal Meteorological Society*, *143*(708), 2665–2676. https://doi.org/10.1002/qj.3115

Verification of uncertain forecasts: A lesson for life?



Good

Confident and correct

Lesson 1

- If you know and you know you know, speak out
- If you don't know and know you don't know, shut up
- If you think you know but actually don't know, please, don't speak out

Lesson 2

Better admit being ignorant than claiming to be confident but half-time correct