Short course SC4.9
Theory and tools of statistical forecast verification

Sebastian Buschow (Bonn), Jochen Bröcker (Reading)

EGU 2023

Outline

Introduction, mathematical setup, and notation

Types of forecasts, scoring rules, and forecast attributes

Tests and p-values

How to cope with dependent data

Verification of spatial fields

Outline

Introduction, mathematical setup, and notation

Types of forecasts, scoring rules, and forecast attributes

Tests and p-values

How to cope with dependent data

Verification of spatial fields

Outline

Introduction, mathematical setup, and notation

Types of forecasts, scoring rules, and forecast attributes

Tests and p-values

How to cope with dependent data

Verification of spatial fields

Outline

Introduction, mathematical setup, and notation

Types of forecasts, scoring rules, and forecast attributes

Tests and p-values

How to cope with dependent data

Verification of spatial fields

Outline

Introduction, mathematical setup, and notation

Types of forecasts, scoring rules, and forecast attributes

Tests and p-values

How to cope with dependent data

Verification of spatial fields

Introduction, mathematical setup, and notation

Types of forecasts, scoring rules, and forecast attributes

Tests and p-values

How to cope with dependent data

Verification of spatial fields

Forecasting systems require verification!

Forecasts (e.g. of meteorological or economical variables) are indispensable for decision support . . . . . . but only if they have a clear statistical interpretation! For instance (proper definition will come later)

Probabilistic forecasts – The forecasts represent the probability distribution of the verification, conditionally on the information available at forecast time.

Mean forecasts – The forecasts represent the conditional mean (expectation value) of the verification.

Quantile forecasts – The forecasts represent a specific conditional quantile of the verification.

Hence, verification has to be in a statistical sense.

Forecasting systems require verification!

Forecasts (e.g. of meteorological or economical variables) are indispensable for decision support . . . . . . but only if they have a clear statistical interpretation! For instance (proper definition will come later)

Probabilistic forecasts – The forecasts represent the probability distribution of the verification, conditionally on the information available at forecast time.

Mean forecasts – The forecasts represent the conditional mean (expectation value) of the verification.

Quantile forecasts – The forecasts represent a specific conditional quantile of the verification.

Hence, verification has to be in a statistical sense.

Forecasting systems require verification!

Forecasts (e.g. of meteorological or economical variables) are indispensable for decision support . . . . . . but only if they have a clear statistical interpretation! For instance (proper definition will come later)

Probabilistic forecasts – The forecasts represent the probability distribution of the verification, conditionally on the information available at forecast time.

Mean forecasts – The forecasts represent the conditional mean (expectation value) of the verification.

Quantile forecasts – The forecasts represent a specific conditional quantile of the verification.

Hence, verification has to be in a statistical sense.

Forecasting systems require verification!

Forecasts (e.g. of meteorological or economical variables) are indispensable for decision support . . . . . . but only if they have a clear statistical interpretation! For instance (proper definition will come later)

Probabilistic forecasts – The forecasts represent the probability distribution of the verification, conditionally on the information available at forecast time.

Mean forecasts – The forecasts represent the conditional mean (expectation value) of the verification.

Quantile forecasts – The forecasts represent a specific conditional quantile of the verification.

Hence, verification has to be in a statistical sense.

Forecasting systems require verification!

Forecasts (e.g. of meteorological or economical variables) are indispensable for decision support . . . . . . but only if they have a clear statistical interpretation! For instance (proper definition will come later)

Probabilistic forecasts – The forecasts represent the probability distribution of the verification, conditionally on the information available at forecast time.

Mean forecasts – The forecasts represent the conditional mean (expectation value) of the verification.

Quantile forecasts – The forecasts represent a specific conditional quantile of the verification.

Hence, verification has to be in a statistical sense.

Forecasting systems require verification!

Forecast verification uses verification–forecast data sets to answer questions regarding "average" forecast behaviour, for instance:

1. Is the proposed statistical interpretation consistent with the actual statistical behaviour (*calibration* or *reliability*)?
2. How much "information" about the verification do the forecasts contain (e.g. when compared to a benchmark forecast) (*resolution*)?

According to the *prequential principles* [Dawid(1984)], forecast verification should *only* take into account

1. the forecasts that have actually been issued,
2. the verifications that have actually materialised,
3. the statistical interpretation of the forecasts, as stated by the forecaster.

Forecasting systems require verification!

Forecast verification uses verification–forecast data sets to answer questions regarding "average" forecast behaviour, for instance:

1. Is the proposed statistical interpretation consistent with the actual statistical behaviour (*calibration* or *reliability*)?
2. How much "information" about the verification do the forecasts contain (e.g. when compared to a benchmark forecast) (*resolution*)?

According to the *prequential principles* [Dawid(1984)], forecast verification should *only* take into account

1. the forecasts that have actually been issued,
2. the verifications that have actually materialised,
3. the statistical interpretation of the forecasts, as stated by the forecaster.

Forecasting systems require verification!

Forecast verification uses verification–forecast data sets to answer questions regarding "average" forecast behaviour, for instance:

1. Is the proposed statistical interpretation consistent with the actual statistical behaviour (*calibration* or *reliability*)?
2. How much "information" about the verification do the forecasts contain (e.g. when compared to a benchmark forecast) (*resolution*)?

According to the *prequential principles* [Dawid(1984)], forecast verification should *only* take into account

1. the forecasts that have actually been issued,
2. the verifications that have actually materialised,
3. the statistical interpretation of the forecasts, as stated by the forecaster.

Forecasting systems require verification!

Forecast verification uses verification–forecast data sets to answer questions regarding "average" forecast behaviour, for instance:

1. Is the proposed statistical interpretation consistent with the actual statistical behaviour (*calibration* or *reliability*)?
2. How much "information" about the verification do the forecasts contain (e.g. when compared to a benchmark forecast) (*resolution*)?

According to the *prequential principles* [Dawid(1984)], forecast verification should *only* take into account

1. the forecasts that have actually been issued,
2. the verifications that have actually materialised,
3. the statistical interpretation of the forecasts, as stated by the forecaster.

Forecasting systems require verification!

Forecast verification uses verification–forecast data sets to answer questions regarding "average" forecast behaviour, for instance:

1. Is the proposed statistical interpretation consistent with the actual statistical behaviour (*calibration* or *reliability*)?
2. How much "information" about the verification do the forecasts contain (e.g. when compared to a benchmark forecast) (*resolution*)?

According to the *prequential principles* [Dawid(1984)], forecast verification should *only* take into account

1. the forecasts that have actually been issued,
2. the verifications that have actually materialised,
3. the statistical interpretation of the forecasts, as stated by the forecaster.

Mathematical setup and notation

Indices in round brackets (.) denote time, subscripts denote vector components (or ensemble members). We are given

- the *verifications* $\{Y(n)\}_{n=1,2,\dots}$, a series of random variables with values in E,
- *state space* E is typically either finite or some subset of $\mathbb{R}^d$.
- the *forecasts* $\{f(n)\}_{n=1,2,\dots}$, a series of random variables with values in some F (depending on type of forecasts),

Note: Forecast $f(n)$ corresponds to verification $Y(n)$; the index n refers to *verification time*. Typically, the forecast $f(n)$ is issued at time $n - L$ where L is the *lead time*.

Mathematical setup and notation

Indices in round brackets (.) denote time, subscripts denote vector components (or ensemble members). We are given

- the *verifications* $\{Y(n)\}_{n=1,2,\ldots}$, a series of random variables with values in E,
- *state space* E is typically either finite or some subset of $\mathbb{R}^d$.
- the *forecasts* $\{f(n)\}_{n=1,2,\ldots}$, a series of random variables with values in some F (depending on type of forecasts),

Note: Forecast $f(n)$ corresponds to verification $Y(n)$; the index n refers to *verification time*. Typically, the forecast $f(n)$ is issued at time $n - L$ where L is the *lead time*.

Mathematical setup and notation

Indices in round brackets (.) denote time, subscripts denote vector components (or ensemble members). We are given

- the *verifications* $\{Y(n)\}_{n=1,2,\ldots}$, a series of random variables with values in E,
- *state space* E is typically either finite or some subset of $\mathbb{R}^d$.
- the *forecasts* $\{f(n)\}_{n=1,2,\ldots}$, a series of random variables with values in some F (depending on type of forecasts),

Note: Forecast $f(n)$ corresponds to verification $Y(n)$; the index n refers to *verification time*. Typically, the forecast $f(n)$ is issued at time $n - L$ where L is the *lead time*.

Mathematical setup and notation

Indices in round brackets (.) denote time, subscripts denote vector components (or ensemble members). We are given

- the *verifications* $\{Y(n)\}_{n=1,2,\ldots}$, a series of random variables with values in E,
- *state space* E is typically either finite or some subset of $\mathbb{R}^d$.
- the *forecasts* $\{f(n)\}_{n=1,2,\ldots}$, a series of random variables with values in some F (depending on type of forecasts),

Note: Forecast $f(n)$ corresponds to verification $Y(n)$; the index n refers to *verification time*. Typically, the forecast $f(n)$ is issued at time $n-L$ where L is the *lead time*.

Mathematical setup and notation

Indices in round brackets (.) denote time, subscripts denote vector components (or ensemble members). We are given

- the *verifications* $\{Y(n)\}_{n=1,2,\ldots}$, a series of random variables with values in E,
- *state space* E is typically either finite or some subset of $\mathbb{R}^d$.
- the *forecasts* $\{f(n)\}_{n=1,2,\ldots}$, a series of random variables with values in some F (depending on type of forecasts),

Note: Forecast $f(n)$ corresponds to verification $Y(n)$; the index n refers to *verification time*. Typically, the forecast $f(n)$ is issued at time $n - L$ where L is the *lead time*.

Mathematical setup and notation

Indices in round brackets (.) denote time, subscripts denote vector components (or ensemble members). We are given

- the *verifications* $\{Y(n)\}_{n=1,2,\ldots}$, a series of random variables with values in E,
- *state space* E is typically either finite or some subset of $\mathbb{R}^d$.
- the *forecasts* $\{f(n)\}_{n=1,2,\ldots}$, a series of random variables with values in some F (depending on type of forecasts),

Note: Forecast $f(n)$ corresponds to verification $Y(n)$; the index n refers to *verification time*. Typically, the forecast $f(n)$ is issued at time $n - L$ where L is the *lead time*.

Types of forecasts

Forecasts can be distinguished based on

Statistical interpretation – Mean, quantile, expectile, cumulative distribution functions, ensembles, . . .

Type of verification – categorical, real, multidimensional, spatial, temporal duration (e.g. of droughts), . . .

Lead time – short range, medium range, seasonal, . . .

We will *not* cover the entire spectrum. Aim is to explain a few core ideas through typical examples. For each example we discuss

- What is the statistical interpretation of the forecast?
- What are desirable forecast attributes?
- How to we verify/quantify these attributes?

Types of forecasts

Forecasts can be distinguished based on

Statistical interpretation – Mean, quantile, expectile, cumulative distribution functions, ensembles, ...

Type of verification – categorical, real, multidimensional, spatial, temporal duration (e.g. of droughts), ...

Lead time – short range, medium range, seasonal, ...

We will *not* cover the entire spectrum. Aim is to explain a few core ideas through typical examples. For each example we discuss

- What is the statistical interpretation of the forecast?
- What are desirable forecast attributes?
- How to we verify/quantify these attributes?

Types of forecasts

Forecasts can be distinguished based on

Statistical interpretation – Mean, quantile, expectile, cumulative distribution functions, ensembles, ...

Type of verification – categorical, real, multidimensional, spatial, temporal duration (e.g. of droughts), ...

Lead time – short range, medium range, seasonal, ...

We will *not* cover the entire spectrum. Aim is to explain a few core ideas through typical examples. For each example we discuss

- What is the statistical interpretation of the forecast?
- What are desirable forecast attributes?
- How to we verify/quantify these attributes?

Types of forecasts

Forecasts can be distinguished based on

Statistical interpretation – Mean, quantile, expectile, cumulative distribution functions, ensembles, . . .

Type of verification – categorical, real, multidimensional, spatial, temporal duration (e.g. of droughts), . . .

Lead time – short range, medium range, seasonal, . . .

We will *not* cover the entire spectrum. Aim is to explain a few core ideas through typical examples. For each example we discuss

- What is the statistical interpretation of the forecast?
- What are desirable forecast attributes?
- How to we verify/quantify these attributes?

Types of forecasts

Forecasts can be distinguished based on

Statistical interpretation – Mean, quantile, expectile, cumulative distribution functions, ensembles, ...

Type of verification – categorical, real, multidimensional, spatial, temporal duration (e.g. of droughts), ...

Lead time – short range, medium range, seasonal, ...

We will *not* cover the entire spectrum. Aim is to explain a few core ideas through typical examples. For each example we discuss

- What is the statistical interpretation of the forecast?
- What are desirable forecast attributes?
- How to we verify/quantify these attributes?

Types of forecasts

Forecasts can be distinguished based on

Statistical interpretation – Mean, quantile, expectile, cumulative distribution functions, ensembles, ...

Type of verification – categorical, real, multidimensional, spatial, temporal duration (e.g. of droughts), ...

Lead time – short range, medium range, seasonal, ...

We will *not* cover the entire spectrum. Aim is to explain a few core ideas through typical examples. For each example we discuss

- What is the statistical interpretation of the forecast?
- What are desirable forecast attributes?
- How to we verify/quantify these attributes?

Probability forecasts

...have a long history!

Probabilité de la pluie suivant la hauteur barométrique. — Elle a été déterminée pour les diverses valeurs de la pression, de 5^{mm} en 5^{mm}. A cet effet, on a compté comme étant de 755^{mm}, toutes les pressions comprises entre $752^{mm},5$ et $757^{mm},5$ et ainsi de suite

Pression.	Nombre de cas de pluie.	Nombre total de cas.	Probab. de pluie.	Pression.	Nombre de cas de pluie.	Nombre total de cas.	Probab. de pluie.
mm				mm			
725.......	1	1	»	755.......	153	293	0,52
730.......	7	11	0,64	760.......	138	325	42
735.......	18	21	86	765.......	89	299	30
740.......	44	60	73	770.......	36	188	19
745.......	88	121	73	775.......	6	26	23
750.......	119	197	60	780.......	.	4	»

(1) Dans le but d'abréger le langage, nous emploierons l'expression de « beau temps » pour indiquer l'absence de pluie entre 9^{h} du matin et minuit.

From [2], observations taken at Parc Montsouris, Paris.

Probability forecasts

Statistical interpretation

Consider probability forecasts for *binary verifications:* $Y(n) \in \{0, 1\}$ for all $n = 1, 2, \ldots$. (Extension to more categories straight forward.) Forecasts $f(n), n = 1, 2, \ldots$ are numbers in $[0, 1]$.

Statistical interpretation:

For $n = 1, 2, \ldots$, the forecast $f(n)$ should be the conditional probability of $Y(n)$ *given* the information available at time $n - L$ (i.e. when the forecast $f(n)$ is issued).

Probability forecasts

Statistical interpretation

Consider probability forecasts for *binary verifications:* $Y(n) \in \{0, 1\}$ for all $n = 1, 2, \ldots$. (Extension to more categories straight forward.) Forecasts $f(n), n = 1, 2, \ldots$ are numbers in $[0, 1]$.

Statistical interpretation:

For $n = 1, 2, \ldots$, the forecast $f(n)$ should be the conditional probability of $Y(n)$ *given* the information available at time $n - L$ (i.e. when the forecast $f(n)$ is issued).

Probability forecasts

Desirable forecast attributes

Specialise our general goals of forecast verification to probability forecasts:

Reliability or Calibration

For probability forecasts, this means

$$f(n) = \mathbb{P}(Y(n) = 1 | f(n)) \quad \text{"Forecast at time } n\text{"} = \text{"Distr. of } Y(n)\text{, given forecast } f(n)\text{"} \tag{1}$$

Resolution

$\mathbb{P}(Y(n) = 1 | f(n))$ exhibits strong variability, i.e. is typically very different from *climatology* $\mathbb{P}(Y(n) = 1)$ [6].

Sharpness

Forecasts $f(n)$ are either close to zero or close to one.

Attention: Only desirable if forecasts are reliable [Gneiting et al.(2007)Gneiting, Balabdaoui, and Raftery]!

Probability forecasts

Desirable forecast attributes

Specialise our general goals of forecast verification to probability forecasts:

Reliability or Calibration

For probability forecasts, this means

$$\begin{aligned} f(n) &= \mathbb{P}(Y(n) = 1 | f(n)) \\ \text{"Forecast at time } n\text{"} &= \text{"Distr. of } Y(n)\text{, given forecast } f(n)\text{"} \end{aligned} \tag{1}$$

Resolution

$\mathbb{P}(Y(n) = 1 | f(n))$ exhibits strong variability, i.e. is typically very different from *climatology* $\mathbb{P}(Y(n) = 1)$ [6].

Sharpness

Forecasts $f(n)$ are either close to zero or close to one.

Attention: Only desirable if forecasts are reliable [Gneiting et al.(2007)Gneiting, Balabdaoui, and Raftery]!

Probability forecasts

Desirable forecast attributes

Specialise our general goals of forecast verification to probability forecasts:

Reliability or Calibration

For probability forecasts, this means

$$\begin{aligned} f(n) &= \mathbb{P}(Y(n) = 1 | f(n)) \\ \text{“Forecast at time } n\text{”} &= \text{“Distr. of } Y(n)\text{, given forecast } f(n)\text{”} \end{aligned} \tag{1}$$

Resolution

$\mathbb{P}(Y(n) = 1|f(n))$ exhibits strong variability, i.e. is typically very different from *climatology* $\mathbb{P}(Y(n) = 1)$ [6].

Sharpness

Forecasts $f(n)$ are either close to zero or close to one.
Attention: Only desirable if forecasts are reliable [Gneiting et al.(2007)Gneiting, Balabdaoui, and Raftery]!

Probability forecasts

Scoring rules

To quantify performance of *individual* forecasts, we use *scoring rules*

$$S(y, f), \qquad \text{where } y = \text{verification}, \ f = \text{forecast}. \qquad (2)$$

Examples for probability forecasts for several categories, i.e. $y \in \{1, \ldots, K\}$ and $f = (f_1, \ldots, f_k)$ with $\sum f_k = 1$:

Logarithmic score – $S(y, f) = \log(f_y)$

Quadratic score – $S(y, f) = \frac{1}{2} \sum_k f_k^2 - f_y$

(CRP score, energy score, ...)

Convention: Smaller score means better forecast.

Probability forecasts

Scoring rules

To quantify performance of *individual* forecasts, we use *scoring rules*

$$S(y, f), \qquad \text{where } y = \text{verification},\ f = \text{forecast}. \tag{2}$$

Examples for probability forecasts for several categories, i.e. $y \in \{1, \ldots, K\}$ and $f = (f_1, \ldots, f_k)$ with $\sum f_k = 1$:

Logarithmic score – $S(y, f) = \log(f_y)$

Quadratic score – $S(y, f) = \frac{1}{2} \sum_k f_k^2 - f_y$

(CRP score, energy score, ...)

Convention: Smaller score means better forecast.

Probability forecasts

Scoring rules

To quantify performance of *individual* forecasts, we use *scoring rules*

$$S(y, f), \qquad \text{where } y = \text{verification}, \; f = \text{forecast}. \tag{2}$$

Examples for probability forecasts for several categories, i.e. $y \in \{1, \ldots, K\}$ and $f = (f_1, \ldots, f_k)$ with $\sum f_k = 1$:

Logarithmic score – $S(y, f) = \log(f_y)$

Quadratic score – $S(y, f) = \frac{1}{2} \sum_k f_k^2 - f_y$

(CRP score, energy score, ...)

Convention: Smaller score means better forecast.

Probability forecasts

Scoring rules

To quantify performance of *individual* forecasts, we use *scoring rules*

$$S(y,f), \qquad \text{where } y = \text{verification}, \; f = \text{forecast}. \tag{2}$$

Examples for probability forecasts for several categories, i.e. $y \in \{1, \dots, K\}$ and $f = (f_1, \dots, f_k)$ with $\sum f_k = 1$:

Logarithmic score – $S(y,f) = \log(f_y)$

Quadratic score – $S(y,f) = \frac{1}{2}\sum_k f_k^2 - f_y$

(CRP score, energy score, ...)

Convention: Smaller score means better forecast.

Probability forecasts

Scoring rules

The mentioned scores are *(strictly) proper scoring rules:* For any probability forecasts f, g we have

$$\sum_k S(k, f) g_k \geq \sum_k S(k, g) g_k \tag{3}$$

i.e. assuming g is correct distribution of Y,

expected score of $f \geq$ expected score of g itself.

We have [3]: $\mathbb{E} S(Y(n), f(n)) = \text{UNC} - \text{RES} + \text{REL}$, (note signs) where for strictly proper scoring rules

- UNC depends only on $Y(n)$, not on $f(n)$,
- RES is positive, unless $f(n)$ has *no* resolution,
- REL is positive, unless $f(n)$ is calibrated.

$$\mathbb{E} S(Y(n), f(n)) \overset{\text{est.}}{\approx} \frac{1}{N} \sum_{n=1}^{N} S(Y(n), f(n))$$

but separate estimates of RES, REL are more difficult to obtain [5].

Probability forecasts

Scoring rules

The mentioned scores are *(strictly) proper scoring rules:* For any probability forecasts f, g we have

$$\sum_k S(k, f) g_k \geq \sum_k S(k, g) g_k \tag{3}$$

i.e. assuming g is correct distribution of Y,

expected score of $f \geq$ expected score of g itself.

We have [3]: $\mathbb{E} S(Y(n), f(n)) = \text{UNC} - \text{RES} + \text{REL}$, (note signs) where for strictly proper scoring rules

- UNC depends only on $Y(n)$, not on $f(n)$,
- RES is positive, unless $f(n)$ has *no* resolution,
- REL is positive, unless $f(n)$ is calibrated.

$$\mathbb{E} S(Y(n), f(n)) \overset{\text{est.}}{\approx} \frac{1}{N} \sum_{n=1}^{N} S(Y(n), f(n))$$

but separate estimates of RES, REL are more difficult to obtain [5].

Probability forecasts

Scoring rules

The mentioned scores are *(strictly) proper scoring rules:* For any probability forecasts f, g we have

$$\sum_k S(k, f) g_k \geq \sum_k S(k, g) g_k \tag{3}$$

i.e. assuming g is correct distribution of Y,

expected score of $f \geq$ expected score of g itself.

We have [3]: $\mathbb{E}S(Y(n), f(n)) = \text{UNC} - \text{RES} + \text{REL}$, (note signs) where for strictly proper scoring rules

- UNC depends only on $Y(n)$, not on $f(n)$,
- RES is positive, unless $f(n)$ has *no* resolution,
- REL is positive, unless $f(n)$ is calibrated.

$$\mathbb{E}S(Y(n), f(n)) \overset{\text{est.}}{\approx} \frac{1}{N} \sum_{n=1}^{N} S(Y(n), f(n))$$

but separate estimates of RES, REL are more difficult to obtain [5].

Probability forecasts

Scoring rules

The mentioned scores are *(strictly) proper scoring rules:* For any probability forecasts f, g we have

$$\sum_k S(k, f) g_k \geq \sum_k S(k, g) g_k \tag{3}$$

i.e. assuming g is correct distribution of Y,

expected score of $f \geq$ expected score of g itself.

We have [3]: $\mathbb{E}S(Y(n), f(n)) = \mathrm{UNC} - \mathrm{RES} + \mathrm{REL}$, (note signs) where for strictly proper scoring rules

- UNC depends only on $Y(n)$, not on $f(n)$,
- RES is positive, unless $f(n)$ has *no* resolution,
- REL is positive, unless $f(n)$ is calibrated.

$$\mathbb{E}S(Y(n), f(n)) \overset{\text{est.}}{\approx} \frac{1}{N} \sum_{n=1}^{N} S(Y(n), f(n))$$

but separate estimates of RES, REL are more difficult to obtain [5].

Probability forecasts

Scoring rules

The mentioned scores are *(strictly) proper scoring rules:* For any probability forecasts f, g we have

$$\sum_k S(k, f) g_k \geq \sum_k S(k, g) g_k \tag{3}$$

i.e. assuming g is correct distribution of Y,

expected score of $f \geq$ expected score of g itself.

We have [3]: $\mathbb{E}S(Y(n), f(n)) = \text{UNC} - \text{RES} + \text{REL}$, (note signs) where for strictly proper scoring rules

- UNC depends only on $Y(n)$, not on $f(n)$,
- RES is positive, unless $f(n)$ has *no* resolution,
- REL is positive, unless $f(n)$ is calibrated.

$$\mathbb{E}S(Y(n), f(n)) \overset{\text{est.}}{\approx} \frac{1}{N} \sum_{n=1}^{N} S(Y(n), f(n))$$

but separate estimates of RES, REL are more difficult to obtain [5].

Probability forecasts
Scoring rules

The mentioned scores are *(strictly) proper scoring rules:* For any probability forecasts f, g we have

$$\sum_k S(k, f) g_k \geq \sum_k S(k, g) g_k \tag{3}$$

i.e. assuming g is correct distribution of Y,

expected score of $f \geq$ expected score of g itself.

We have [3]: $\mathbb{E}S(Y(n), f(n)) = \text{UNC} - \text{RES} + \text{REL}$, (note signs) where for strictly proper scoring rules

- UNC depends only on $Y(n)$, not on $f(n)$,
- RES is positive, unless $f(n)$ has *no* resolution,
- REL is positive, unless $f(n)$ is calibrated.

$$\mathbb{E}S(Y(n), f(n)) \overset{\text{est.}}{\approx} \frac{1}{N} \sum_{n=1}^{N} S(Y(n), f(n))$$

but separate estimates of RES, REL are more difficult to obtain [5].

Probability forecasts

Identification functions

More relevant for testing are *identification functions*, i.e. functions $\mathsf{V}(y, f)$ [13, 9] so that

$$\sum_k \mathsf{V}(k, f) f_k = 0 \tag{4}$$

(but typically $\sum_k \mathsf{V}(k, g) f_k \neq 0$ if $f \neq g$). Note: V can be multi–dimensional.

Example: Identification function V with components $V_d(y, f) = \mathbb{1}_{\{y=d\}} - f_d$ for $d = 1, \ldots, K$. For *Conditional mean* forecasts or binary probability forecasts take $\mathsf{V}(f, y) = f - y$.

If forecasts are reliable ...

$$\mathbb{E}(\mathsf{V}(Y(n), f(n)) | f(n)) = 0 \qquad \text{for } n = 1, 2, \ldots. \tag{5}$$

Probability forecasts

Identification functions

More relevant for testing are *identification functions*, i.e. functions $\mathrm{V}(y, f)$ [13, 9] so that

$$\sum_k \mathrm{V}(k, f) f_k = 0 \tag{4}$$

(but typically $\sum_k \mathrm{V}(k, g) f_k \neq 0$ if $f \neq g$). Note: V can be multi–dimensional.

Example: Identification function V with components $V_d(y, f) = \mathbb{1}_{\{y=d\}} - f_d$ for $d = 1, \ldots, K$. For *Conditional mean* forecasts or binary probability forecasts take $\mathrm{V}(f, y) = f - y$.

If forecasts are reliable ...

$$\mathbb{E}(\mathrm{V}(Y(n), f(n)) | f(n)) = 0 \qquad \text{for } n = 1, 2, \ldots. \tag{5}$$

Probability forecasts

Identification functions

More relevant for testing are *identification functions*, i.e. functions $\mathrm{V}(y, f)$ [13, 9] so that

$$\sum_k \mathrm{V}(k, f) f_k = 0 \tag{4}$$

(but typically $\sum_k \mathrm{V}(k, g) f_k \neq 0$ if $f \neq g$). Note: V can be multi–dimensional.

Example: Identification function V with components $V_d(y, f) = \mathbb{1}_{\{y=d\}} - f_d$ for $d = 1, \ldots, K$. For *Conditional mean* forecasts or binary probability forecasts take $\mathrm{V}(f, y) = f - y$.

If forecasts are reliable ...

$$\mathbb{E}(\mathrm{V}(Y(n), f(n)) | f(n)) = 0 \qquad \text{for } n = 1, 2, \ldots. \tag{5}$$

Probability forecasts

Identification functions

More relevant for testing are *identification functions*, i.e. functions $\mathsf{V}(y, f)$ [13, 9] so that

$$\sum_k \mathsf{V}(k, f) f_k = 0 \tag{4}$$

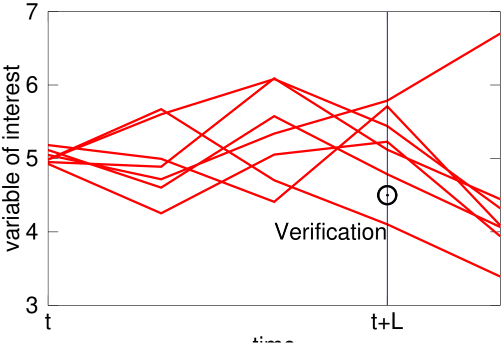
(but typically $\sum_k \mathsf{V}(k, g) f_k \neq 0$ if $f \neq g$). Note: V can be multi–dimensional.

Example: Identification function V with components $V_d(y, f) = \mathbb{1}_{\{y=d\}} - f_d$ for $d = 1, \ldots, K$. For *Conditional mean* forecasts or binary probability forecasts take $\mathsf{V}(f, y) = f - y$.

If forecasts are reliable ...

$$\mathbb{E}(\mathsf{V}(Y(n), f(n)) | f(n)) = 0 \qquad \text{for } n = 1, 2, \ldots. \tag{5}$$

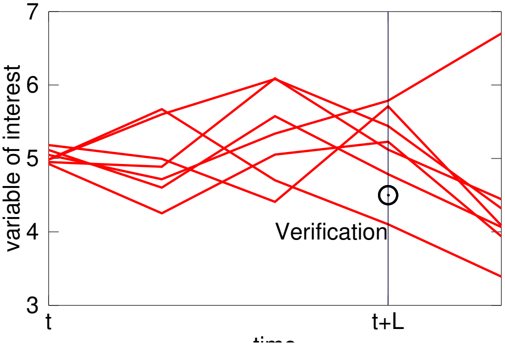
Ensemble forecasts



Meteorological *ensemble forecasts* typically are numerical simulations of the atmosphere with heterogenous initial conditions (obtained using *data assimilation*).

Notation: Let $X(n) = (X_1(n), \ldots, X_{K-1}(n))$, $n = 1, 2, \ldots$ ensemble forecasts. Let $\mathcal{F}(n)$, $n = 1, 2, \ldots$ be information available to forecaster at time $n - L$.

Ensemble forecasts



Meteorological *ensemble forecasts* typically are numerical simulations of the atmosphere with heterogenous initial conditions (obtained using *data assimilation*).

Notation: Let $X(n) = (X_1(n), \ldots, X_{K-1}(n)), n = 1, 2, \ldots$ ensemble forecasts. Let $\mathcal{F}(n), n = 1, 2, \ldots$ be information available to forecaster at time $n - L$.

Ensemble forecasts

Statistical interpretation and desirable forecast attributes

Reliability

Ensemble members $(X_1(n), \ldots, X_{K-1}(n))$ and verification $Y(n)$ should be "independent draws" from the conditional distribution $\mathbb{P}(Y(N)|\mathcal{F}(n))$ [3, 4, 15].

Resolution

If $Y(m)$ and $Y(n)$ are very different for $m \neq n$, then $\mathrm{X}(m)$ and $\mathrm{X}(n)$ should also be very different.

Sharpness

Ensembles $\mathrm{X}(n)$ is very "narrow" (i.e. little spread). Attention: Only desirable for reliable ensembles.

Ensemble forecasts

Statistical interpretation and desirable forecast attributes

Reliability

Ensemble members $(X_1(n), \ldots, X_{K-1}(n))$ and verification $Y(n)$ should be "independent draws" from the conditional distribution $\mathbb{P}(Y(N)|\mathcal{F}(n))$ [3, 4, 15].

Resolution

If $Y(m)$ and $Y(n)$ are very different for $m \neq n$, then $\mathrm{X}(m)$ and $\mathrm{X}(n)$ should also be very different.

Sharpness

Ensembles $\mathrm{X}(n)$ is very "narrow" (i.e. little spread). Attention: Only desirable for reliable ensembles.

Ensemble forecasts

Statistical interpretation and desirable forecast attributes

Reliability

Ensemble members $(X_1(n), \ldots, X_{K-1}(n))$ and verification $Y(n)$ should be "independent draws" from the conditional distribution $\mathbb{P}(Y(N)|\mathcal{F}(n))$ [3, 4, 15].

Resolution

If $Y(m)$ and $Y(n)$ are very different for $m \neq n$, then $\mathrm{X}(m)$ and $\mathrm{X}(n)$ should also be very different.

Sharpness

Ensembles $\mathrm{X}(n)$ is very "narrow" (i.e. little spread). Attention: Only desirable for reliable ensembles.

Ensemble forecasts

Scoring rules and rank histograms

We assume $E = \mathbb{R}$ for simplicity.
Scoring rules: The CRP Score can directly be applied to ensembles. Other scores require postprocessing (e.g. kernel methods)
The rank histogram: Define $R(n)$ to be the rank of $Y(n)$ among the ensemble members $X_1(n), \ldots, X_{K-1}(n)$. Reliability (plus technical conditions) implies

$$\mathbb{P}(R(n) = k) = \frac{1}{K}.$$

In particular, the ranks have a uniform distribution.
Uniform rank distribution has long been recognized as necessary consequence of reliability [1, 14, 11, 10]. Can be cast in terms of identification functions.

Ensemble forecasts

Scoring rules and rank histograms

We assume $E = \mathbb{R}$ for simplicity.
Scoring rules: The CRP Score can directly be applied to ensembles. Other scores require postprocessing (e.g. kernel methods)
The rank histogram: Define $R(n)$ to be the rank of $Y(n)$ among the ensemble members $X_1(n), \ldots, X_{K-1}(n)$. Reliability (plus technical conditions) implies

$$\mathbb{P}(R(n) = k) = \frac{1}{K}.$$

In particular, the ranks have a uniform distribution.
Uniform rank distribution has long been recognized as necessary consequence of reliability [1, 14, 11, 10]. Can be cast in terms of identification functions.

Ensemble forecasts

Scoring rules and rank histograms

We assume $E = \mathbb{R}$ for simplicity.
Scoring rules: The CRP Score can directly be applied to ensembles. Other scores require postprocessing (e.g. kernel methods)
The rank histogram: Define $R(n)$ to be the rank of $Y(n)$ among the ensemble members $X_1(n), \ldots, X_{K-1}(n)$. Reliability (plus technical conditions) implies

$$\mathbb{P}(R(n) = k) = \frac{1}{K}.$$

In particular, the ranks have a uniform distribution.
Uniform rank distribution has long been recognized as necessary consequence of reliability [1, 14, 11, 10]. Can be cast in terms of identification functions.

We assume $E = \mathbb{R}$ for simplicity.
Scoring rules: The CRP Score can directly be applied to ensembles. Other scores require postprocessing (e.g. kernel methods)
The rank histogram: Define $R(n)$ to be the rank of $Y(n)$ among the ensemble members $X_1(n), \ldots, X_{K-1}(n)$. Reliability (plus technical conditions) implies

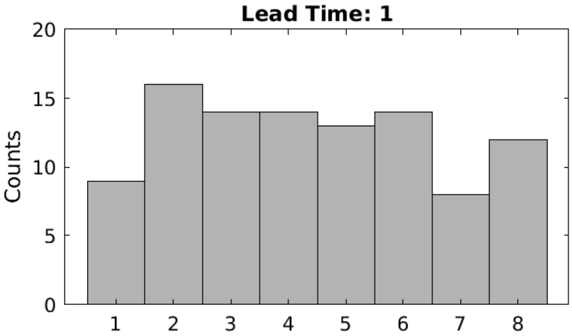
$$\mathbb{P}(R(n) = k) = \frac{1}{K}.$$

In particular, the ranks have a uniform distribution.
Uniform rank distribution has long been recognized as necessary consequence of reliability [1, 14, 11, 10]. Can be cast in terms of identification functions.

Ensemble forecasts

Rank histograms

Qualitative measure of reliability is the *rank histogram*



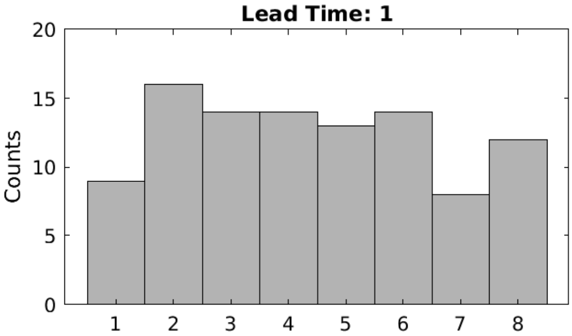
Quantitative tests encounter two problems:

1. ranks are serially correlated,
2. uniform histogram only necessary but not sufficient for reliability.

Ensemble forecasts

Rank histograms

Qualitative measure of reliability is the *rank histogram*



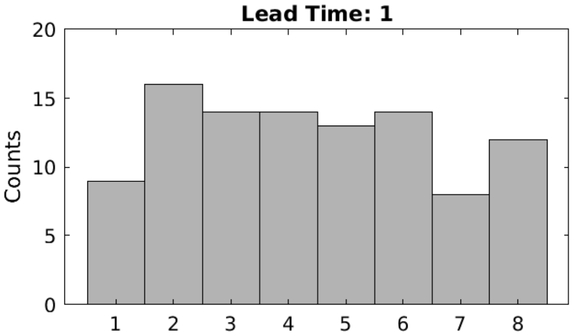
Quantitative tests encounter two problems:

1. ranks are serially correlated,
2. uniform histogram only necessary but not sufficient for reliability.

Ensemble forecasts

Rank histograms

Qualitative measure of reliability is the *rank histogram*



Quantitative tests encounter two problems:

1. ranks are serially correlated,
2. uniform histogram only necessary but not sufficient for reliability.

What is a "statistical test" for reliability?

The verification–forecast pairs $\{(Y(n), f(n)), n = 1, 2, \ldots\}$ form a series of random variables with joint distribution $\mathbb{P}$. Reliability imposes a constraint on $\mathbb{P}$. Write

$\mathcal{H}_0$ *Null Hypothesis:* – all distributions $\mathbb{P}$ that satisfy reliability,

$\mathcal{H}_1$ *Alternative:* – a set of distributions that do *not* satisfy reliability (might be the complement of $\mathcal{H}_0$).

A *test statistic* τ is a function of the data $\{(Y(n), f(n))\}_{n \leq N}$.

Goal of testing

Find a test statistic τ and threshold c so that $\mathbb{P}(\tau \geq c)$ is small if $\mathbb{P} \in \mathcal{H}_0$, but large if $\mathbb{P} \in \mathcal{H}_1$.

Can be seen as *False Alarm Rate* resp *Hit Rate* of the test "$\tau \geq c$".

What is a "statistical test" for reliability?

The verification–forecast pairs $\{(Y(n), f(n)), n = 1, 2, \ldots\}$ form a series of random variables with joint distribution $\mathbb{P}$. Reliability imposes a constraint on $\mathbb{P}$. Write

$\mathcal{H}_0$ *Null Hypothesis:* – all distributions $\mathbb{P}$ that satisfy reliability,

$\mathcal{H}_1$ *Alternative:* – a set of distributions that do *not* satisfy reliability (might be the complement of $\mathcal{H}_0$).

A *test statistic* τ is a function of the data $\{(Y(n), f(n))\}_{n \leq N}$.

Goal of testing

Find a test statistic τ and threshold c so that $\mathbb{P}(\tau \geq c)$ is small if $\mathbb{P} \in \mathcal{H}_0$, but large if $\mathbb{P} \in \mathcal{H}_1$.

Can be seen as *False Alarm Rate* resp *Hit Rate* of the test "$\tau \geq c$".

What is a "statistical test" for reliability?

The verification–forecast pairs $\{(Y(n), f(n)), n = 1, 2, \dots\}$ form a series of random variables with joint distribution $\mathbb{P}$. Reliability imposes a constraint on $\mathbb{P}$. Write

$\mathcal{H}_0$ *Null Hypothesis:* – all distributions $\mathbb{P}$ that satisfy reliability,

$\mathcal{H}_1$ *Alternative:* – a set of distributions that do *not* satisfy reliability (might be the complement of $\mathcal{H}_0$).

A *test statistic* τ is a function of the data $\{(Y(n), f(n))\}_{n \leq N}$.

Goal of testing

Find a test statistic τ and threshold c so that $\mathbb{P}(\tau \geq c)$ is small if $\mathbb{P} \in \mathcal{H}_0$, but large if $\mathbb{P} \in \mathcal{H}_1$.

Can be seen as *False Alarm Rate* resp *Hit Rate* of the test "$\tau \geq c$".

What is a "statistical test" for reliability?

The verification–forecast pairs $\{(Y(n), f(n)), n = 1, 2, \ldots\}$ form a series of random variables with joint distribution $\mathbb{P}$. Reliability imposes a constraint on $\mathbb{P}$. Write

$\mathcal{H}_0$ *Null Hypothesis:* – all distributions $\mathbb{P}$ that satisfy reliability,

$\mathcal{H}_1$ *Alternative:* – a set of distributions that do *not* satisfy reliability (might be the complement of $\mathcal{H}_0$).

A *test statistic* τ is a function of the data $\{(Y(n), f(n))\}_{n \leq N}$.

Goal of testing

Find a test statistic τ and threshold c so that $\mathbb{P}(\tau \geq c)$ is small if $\mathbb{P} \in \mathcal{H}_0$, but large if $\mathbb{P} \in \mathcal{H}_1$.

Can be seen as *False Alarm Rate* resp *Hit Rate* of the test "$\tau \geq c$".

What is a "statistical test" for reliability?

The verification–forecast pairs $\{(Y(n), f(n)), n = 1, 2, \ldots\}$ form a series of random variables with joint distribution $\mathbb{P}$. Reliability imposes a constraint on $\mathbb{P}$. Write

$\mathcal{H}_0$ *Null Hypothesis:* – all distributions $\mathbb{P}$ that satisfy reliability,

$\mathcal{H}_1$ *Alternative:* – a set of distributions that do *not* satisfy reliability (might be the complement of $\mathcal{H}_0$).

A *test statistic* τ is a function of the data $\{(Y(n), f(n))\}_{n \leq N}$.

Goal of testing

Find a test statistic τ and threshold c so that $\mathbb{P}(\tau \geq c)$ is small if $\mathbb{P} \in \mathcal{H}_0$, but large if $\mathbb{P} \in \mathcal{H}_1$.

Can be seen as *False Alarm Rate* resp *Hit Rate* of the test "$\tau \geq c$".

The power function

Important points to keep in mind:

- $\mathbb{P}(\tau \geq c)$ takes different values for different $\mathbb{P} \in \mathcal{H}_0$ so no single "False Alarm rate". Might define

$$\mathrm{FAR} := \max_{\mathbb{P} \in \mathcal{H}_0} \mathbb{P}(\tau \geq c). \tag{6}$$

- Same for "Hit Rate", so define

$$\mathrm{HR} := \min_{\mathbb{P} \in \mathcal{H}_1} \mathbb{P}(\tau \geq c). \tag{7}$$

- Typically, $\mathrm{HR} \gg \mathrm{FAR}$ only for increasing amounts of data.

The power function

Important points to keep in mind:

- $\mathbb{P}(\tau \geq c)$ takes different values for different $\mathbb{P} \in \mathcal{H}_0$ so no single "False Alarm rate". Might define

$$\text{FAR} := \max_{\mathbb{P} \in \mathcal{H}_0} \mathbb{P}(\tau \geq c). \tag{6}$$

- Same for "Hit Rate", so define

$$\text{HR} := \min_{\mathbb{P} \in \mathcal{H}_1} \mathbb{P}(\tau \geq c). \tag{7}$$

- Typically, HR $\gg$ FAR only for increasing amounts of data.

The power function

Important points to keep in mind:

- $\mathbb{P}(\tau \geq c)$ takes different values for different $\mathbb{P} \in \mathcal{H}_0$ so no single "False Alarm rate". Might define

$$\text{FAR} := \max_{\mathbb{P} \in \mathcal{H}_0} \mathbb{P}(\tau \geq c). \tag{6}$$

- Same for "Hit Rate", so define

$$\text{HR} := \min_{\mathbb{P} \in \mathcal{H}_1} \mathbb{P}(\tau \geq c). \tag{7}$$

- Typically, HR $\gg$ FAR only for increasing amounts of data.

How to chose test statistic τ

REL part of score decomposition Taking τ an estimate of *REL* would be expected to be small under $\mathcal{H}_0$ yet large under $\mathcal{H}_1$. For most scores, distribution of τ is classic for independent d$\{(Y(n), f(n))\}$ but otherwise FAR difficult to compute.

Identification function V Since $\mathbb{E}(\mathrm{V}(Y(n), f(n))) = 0$ under $\mathcal{H}_0$,

$$R(N) = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \mathrm{V}(Y(k), f(k))$$

satisfies CLT under fairly general conditions. May then take $\tau = R(N)^{\mathrm{t}} \Gamma^{-1} R(N)$ as test statistic. Issues:

1. Covariance Γ of R_N – needs estimating.
2. Power – will be quite low.

How to chose test statistic τ

REL part of score decomposition Taking τ an estimate of *REL* would be expected to be small under $\mathcal{H}_0$ yet large under $\mathcal{H}_1$. For most scores, distribution of τ is classic for independent $\mathrm{d}\{(Y(n), f(n))\}$ but otherwise FAR difficult to compute.

Identification function V Since $\mathbb{E}(\mathrm{V}(Y(n), f(n))) = 0$ under $\mathcal{H}_0$,

$$R(N) = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \mathrm{V}(Y(k), f(k))$$

satisfies CLT under fairly general conditions. May then take $\tau = R(N)^{\mathrm{t}} \Gamma^{-1} R(N)$ as test statistic. Issues:

1. Covariance Γ of R_N – needs estimating.
2. Power – will be quite low.

How to chose test statistic τ

REL part of score decomposition Taking τ an estimate of *REL* would be expected to be small under $\mathcal{H}_0$ yet large under $\mathcal{H}_1$. For most scores, distribution of τ is classic for independent d$\{(Y(n), f(n))\}$ but otherwise FAR difficult to compute.

Identification function V Since $\mathbb{E}(\mathrm{V}(Y(n), f(n))) = 0$ under $\mathcal{H}_0$,

$$R(N) = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \mathrm{V}(Y(k), f(k))$$

satisfies CLT under fairly general conditions. May then take $\tau = R(N)^{\mathrm{t}} \Gamma^{-1} R(N)$ as test statistic. Issues:

1. Covariance Γ of R_N – needs estimating.
2. Power – will be quite low.

How to chose test statistic τ

REL part of score decomposition Taking τ an estimate of *REL* would be expected to be small under $\mathcal{H}_0$ yet large under $\mathcal{H}_1$. For most scores, distribution of τ is classic for independent d$\{(Y(n), f(n))\}$ but otherwise FAR difficult to compute.

Identification function V Since $\mathbb{E}(\mathrm{V}(Y(n), f(n))) = 0$ under $\mathcal{H}_0$,

$$R(N) = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \mathrm{V}(Y(k), f(k))$$

satisfies CLT under fairly general conditions. May then take $\tau = R(N)^{\mathrm{t}} \Gamma^{-1} R(N)$ as test statistic. Issues:

1. Covariance Γ of R_N – needs estimating.
2. Power – will be quite low.

A stronger reliability condition

To deal with both issues

Recall that $\mathcal{F}(n)$ is the information available to the forecaster at time $n - L$ (i.e. when issuing forecast $f(n)$), for $n = 1, 2, \ldots$. We impose the

Stronger reliability condition

$$\mathbb{P}(Y(n) = k | \mathcal{F}(n)) = f_k(n)$$

i.e. conditioning on *all available information.*

Compare with Eq. (1) where conditioning is only on current forecast. We also impose

Stationarity

Distribution of data independent of time origin.

A stronger reliability condition

To deal with both issues

Recall that $\mathcal{F}(n)$ is the information available to the forecaster at time $n - L$ (i.e. when issuing forecast $f(n)$), for $n = 1, 2, \ldots$. We impose the

Stronger reliability condition

$$\mathbb{P}(Y(n) = k | \mathcal{F}(n)) = f_k(n)$$

i.e. conditioning on *all available information*.

Compare with Eq. (1) where conditioning is only on current forecast. We also impose

Stationarity

Distribution of data independent of time origin.

A stronger reliability condition

To deal with both issues

Recall that $\mathcal{F}(n)$ is the information available to the forecaster at time $n - L$ (i.e. when issuing forecast $f(n)$), for $n = 1, 2, \ldots$. We impose the

Stronger reliability condition

$$\mathbb{P}(Y(n) = k | \mathcal{F}(n)) = f_k(n)$$

i.e. conditioning on *all available information*.

Compare with Eq. (1) where conditioning is only on current forecast. We also impose

Stationarity

Distribution of data independent of time origin.

How to improve power: Stratification

A *stratification* is any process $\{S(n), n \in \mathbb{N}\}$ so that $S(n)$ is "part of" $\mathcal{F}(n)$ for $n = 1, 2 \ldots$.

Examples

1. $S(k) = \text{const}$ (this would test for unconditional reliability)
2. $S(k) = f(k)$ (essentially a regression based test; this often performs very well already)

Strong reliability implies $\mathbb{E}(\mathrm{V}(n)S(n)) = 0$ for all $n \in \mathbb{N}$, thus we consider

$$R(N) = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \mathrm{V}(Y(k), f(k))S(k)$$

and take $\tau = R(N)^{\mathrm{t}} \Gamma^{-1} R(N)$ as test statistic.

How to improve power: Stratification

A *stratification* is any process $\{S(n), n \in \mathbb{N}\}$ so that $S(n)$ is "part of" $\mathcal{F}(n)$ for $n = 1, 2 \ldots$.

Examples

1. $S(k) = \text{const}$ (this would test for unconditional reliability)
2. $S(k) = f(k)$ (essentially a regression based test; this often performs very well already)

Strong reliability implies $\mathbb{E}(\mathrm{V}(n)S(n)) = 0$ for all $n \in \mathbb{N}$, thus we consider

$$R(N) = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \mathrm{V}(Y(k), f(k))S(k)$$

and take $\tau = R(N)^{t}\Gamma^{-1}R(N)$ as test statistic.

How to improve power: Stratification

A *stratification* is any process $\{S(n), n \in \mathbb{N}\}$ so that $S(n)$ is "part of" $\mathcal{F}(n)$ for $n = 1, 2 \ldots$.

Examples

1. $S(k) = \text{const}$ (this would test for unconditional reliability)
2. $S(k) = f(k)$ (essentially a regression based test; this often performs very well already)

Strong reliability implies $\mathbb{E}(V(n)S(n)) = 0$ for all $n \in \mathbb{N}$, thus we consider

$$R(N) = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} V(Y(k), f(k))S(k)$$

and take $\tau = R(N)^{t}\Gamma^{-1}R(N)$ as test statistic.

How to improve power: Stratification

A *stratification* is any process $\{S(n), n \in \mathbb{N}\}$ so that $S(n)$ is "part of" $\mathcal{F}(n)$ for $n = 1, 2 \ldots$.

Examples

1. $S(k) = \text{const}$ (this would test for unconditional reliability)
2. $S(k) = f(k)$ (essentially a regression based test; this often performs very well already)

Strong reliability implies $\mathbb{E}(\mathrm{V}(n)S(n)) = 0$ for all $n \in \mathbb{N}$, thus we consider

$$R(N) = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \mathrm{V}(Y(k), f(k)) S(k)$$

and take $\tau = R(N)^{\mathrm{t}} \Gamma^{-1} R(N)$ as test statistic.

Estimating variance of $R(N)$

For $\tau = R(N)^{\mathrm{t}}\Gamma^{-1}R(N)$, we need and estimate $\hat{\Gamma}$ of $\Gamma = \mathrm{Cov}(R(N))$, which is asymptotically given by

$$\Gamma = \mathrm{Cov}(\psi(0)) + 2\sum_{k} \mathrm{Cov}(\psi(0), \psi(k)) \tag{8}$$

where $\psi(k) := \mathrm{V}(Y(k), f(k))S(k)$. Due to strong reliability assumption, sum has *only L nonzero terms* which we estimate one-by-one.

Theorem [7]

Under $\mathcal{H}_0$, test statistic τ has asymptotic χ^2 distribution.

Estimating variance of $R(N)$

For $\tau = R(N)^{\mathrm{t}}\Gamma^{-1}R(N)$, we need and estimate $\hat{\Gamma}$ of $\Gamma = \mathrm{Cov}(R(N))$, which is asymptotically given by

$$\Gamma = \mathrm{Cov}(\psi(0)) + 2\sum_{k} \mathrm{Cov}(\psi(0), \psi(k)) \tag{8}$$

where $\psi(k) := \mathrm{V}(Y(k), f(k))S(k)$. Due to strong reliability assumption, sum has *only L nonzero terms* which we estimate one-by-one.

Theorem [7]

Under $\mathcal{H}_0$, test statistic τ has asymptotic χ^2 distribution.

Estimating variance of $R(N)$

For $\tau = R(N)^{\mathrm{t}}\Gamma^{-1}R(N)$, we need and estimate $\hat{\Gamma}$ of $\Gamma = \mathrm{Cov}(R(N))$, which is asymptotically given by

$$\Gamma = \mathrm{Cov}(\psi(0)) + 2\sum_k \mathrm{Cov}(\psi(0), \psi(k)) \tag{8}$$

where $\psi(k) := \mathrm{V}(Y(k), f(k))S(k)$. Due to strong reliability assumption, sum has *only L nonzero terms* which we estimate one-by-one.

Theorem [7]

Under $\mathcal{H}_0$, test statistic τ has asymptotic χ^2 distribution.

Stratified rank histograms for ensemble forecasts

Bonus material

A more detailed picture of the reliability can be obtained through *stratified* rank histograms.

Define *strata*

$$S(n) := s(Y(n), X_1(n), \ldots, X_{K-1}(n)),$$

where $s : E^K \to \{1, \ldots, L\}$ is a symmetric function assuming only L different values (e.g. coarse grained empirical mean or median).

Reliability implies that the ranks have a uniform distribution *in each stratum*, i.e.

$$\mathbb{P}(R(n) = k | S(n) = l) = \frac{1}{K} \qquad \text{for all } l = 1, \ldots, L.$$

Stratified rank histograms for ensemble forecasts

Bonus material

A more detailed picture of the reliability can be obtained through *stratified* rank histograms.

Define *strata*

$$S(n) := s(Y(n), X_1(n), \ldots, X_{K-1}(n)),$$

where $s : E^K \to \{1, \ldots, L\}$ is a symmetric function assuming only L different values (e.g. coarse grained empirical mean or median).

Reliability implies that the ranks have a uniform distribution *in each stratum*, i.e.

$$\mathbb{P}(R(n) = k | S(n) = l) = \frac{1}{K} \qquad \text{for all } l = 1, \ldots, L.$$

Stratified rank histograms for ensemble forecasts

A more detailed picture of the reliability can be obtained through *stratified* rank histograms.

Define *strata*

$$S(n) := s(Y(n), X_1(n), \ldots, X_{K-1}(n)),$$

where $s : E^K \to \{1, \ldots, L\}$ is a symmetric function assuming only L different values (e.g. coarse grained empirical mean or median).

Reliability implies that the ranks have a uniform distribution *in each stratum*, i.e.

$$\mathbb{P}(R(n) = k | S(n) = l) = \frac{1}{K} \qquad \text{for all } l = 1, \ldots, L.$$

A generalised GOF test for stratified rank histograms

Bonus material

Consider $N_{k,l} := \#\{n; R(n) = k, S(n) = l\}$ for $k = 1, \dots, K$ and $l = 1, \dots, L$, and set

$$Z_{k,l} := \frac{N_{k,l} - \frac{1}{K}\sum_k N_{k,l}}{\sqrt{K \sum_k N_{k,l}}}$$

Theorem (J.B. et al [8])

Assume all $\pi(n)$ have continuous CDF's and that $\{(R(n), S(n))\}_{n\in\mathbb{N}}$ is ergodic. Then $(Z_{k,l})_{k,l}$ is asymptotically normal with mean zero and some covariance Γ, which has rank $(K-1)L$. Further, there exists a consistent estimator $\hat{\Gamma}^+$ for Γ^+. Hence, the test statistic

$$t := Z^T \hat{\Gamma}^+ Z$$

is asymptotically χ–square with $(K-1)L$ dof.

A generalised GOF test for stratified rank histograms

Bonus material

Consider $N_{k,l} := \#\{n; R(n) = k, S(n) = l\}$ for $k = 1, \ldots, K$ and $l = 1, \ldots, L$, and set

$$Z_{k,l} := \frac{N_{k,l} - \frac{1}{K}\sum_k N_{k,l}}{\sqrt{K \sum_k N_{k,l}}}$$

Theorem (J.B. et al [8])

Assume all $\pi(n)$ have continuous CDF's and that $\{(R(n), S(n))\}_{n\in\mathbb{N}}$ is ergodic. Then $(Z_{k,l})_{k,l}$ is asymptotically normal with mean zero and some covariance Γ, which has rank $(K-1)L$. Further, there exists a consistent estimator $\hat{\Gamma}^+$ for Γ^+. Hence, the test statistic

$$t := Z^T \hat{\Gamma}^+ Z$$

is asymptotically χ–square with $(K-1)L$ dof.

A generalised GOF test for stratified rank histograms

Bonus material

Consider $N_{k,l} := \#\{n; R(n) = k, S(n) = l\}$ for $k = 1, \ldots, K$ and $l = 1, \ldots, L$, and set

$$Z_{k,l} := \frac{N_{k,l} - \frac{1}{K}\sum_k N_{k,l}}{\sqrt{K \sum_k N_{k,l}}}$$

Theorem (J.B. et al [8])

Assume all $\pi(n)$ have continuous CDF's and that $\{(R(n), S(n))\}_{n \in \mathbb{N}}$ is ergodic. Then $(Z_{k,l})_{k,l}$ is asymptotically normal with mean zero and some covariance Γ, which has rank $(K-1)L$. Further, there exists a consistent estimator $\hat{\Gamma}^+$ for Γ^+. Hence, the test statistic

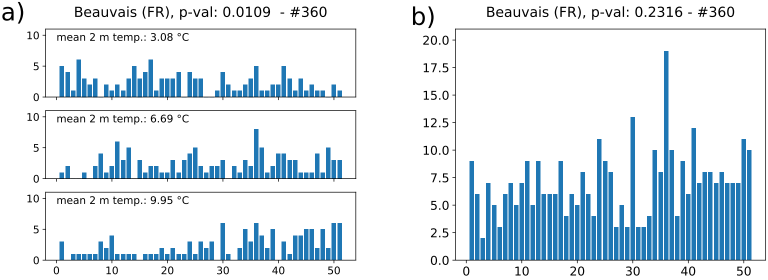
$$t := Z^T \hat{\Gamma}^+ Z$$

is asymptotically χ–square with $(K-1)L$ dof.

Num. Example: ECMWF temperature data

Bonus material

Verification $\{Y(n)\}$ are daily two-metre temperature observations in Beauvais. Ensembles come from the ECMWF operational medium range ensemble prediction system (lead time 5 days). Forecasts were classified into three *strata* corresponding to warm, medium, and cold situations.

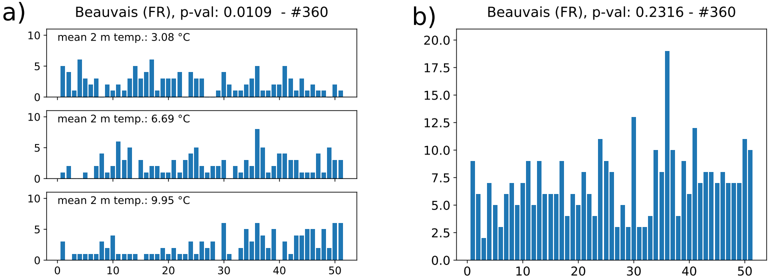


Unstratified histogram (right panel b) shows no evidence for lack of reliability but stratification (left panel a) reveals significant conditional bias (forecast is too warm in cold conditions and too cold in warm conditions), leading to lack of reliability.

Num. Example: ECMWF temperature data

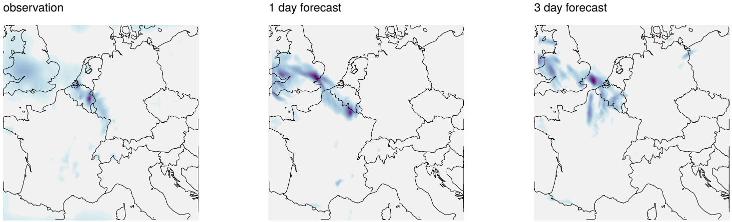
Bonus material

Verification $\{Y(n)\}$ are daily two-metre temperature observations in Beauvais. Ensembles come from the ECMWF operational medium range ensemble prediction system (lead time 5 days). Forecasts were classified into three *strata* corresponding to warm, medium, and cold situations.



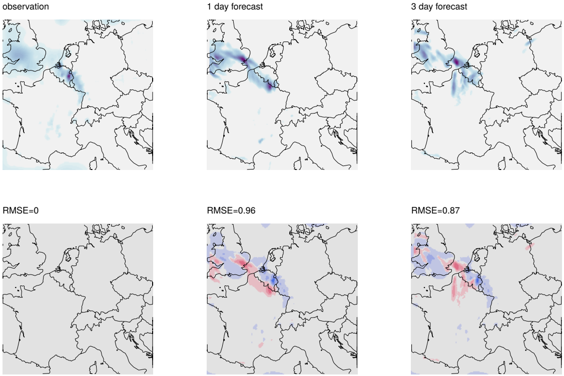
Unstratified histogram (right panel b) shows no evidence for lack of reliability but stratification (left panel a) reveals significant conditional bias (forecast is too warm in cold conditions and too cold in warm conditions), leading to lack of reliability.

Which forecast field is best?



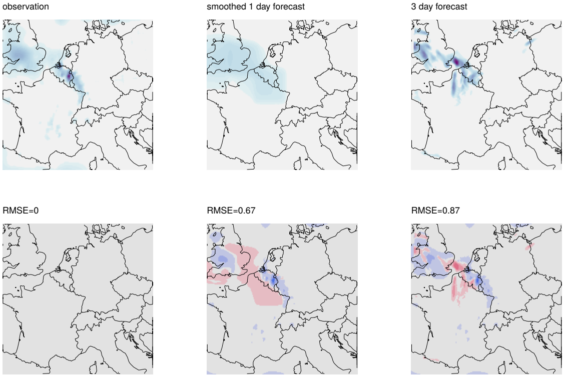
(MesoVICT standardised test case 2007-07-20 11 UTC, BOLAM hourly precipitation forecasts vs station-based VERA reference field.)

Which forecast field is best?



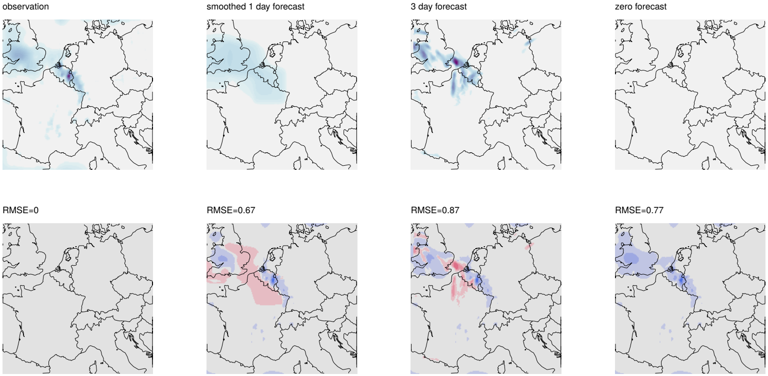
Double penalty: Mis-placed features punished twice, grid-point wise verification not helpful!

Which forecast field is best?



Double penalty: Mis-placed features punished twice, grid-point wise verification not helpful!

Which forecast field is best?

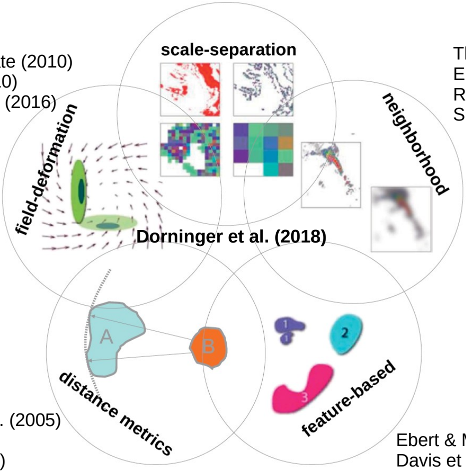


Double penalty: Mis-placed features punished twice, grid-point wise verification not helpful!

Briggs & Levine (1997)
Casati et al. (2004)
Yano & Jakubiak (2016)
Kapp et al. (2018)

Keil & Craig (2007)
Marzban & Sandgate (2010)
Gilleland et al. (2010)
Han and Szunyogh (2016)
Farchi et al. (2016)

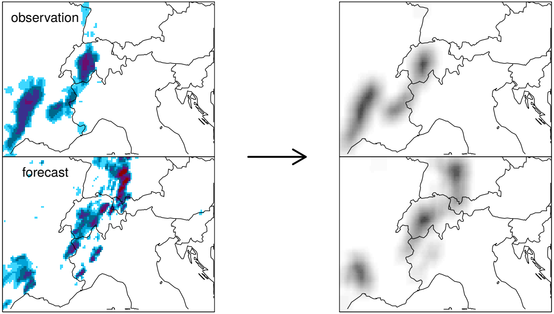
Theis et al. (2005)
Ebert (2008)
Roberts & Lean (2008)
Skok & Hladnik (2018)



Venugopal et al. (2005)
Gilleland (2011)
Zhu et al. (2011)
Gilleland (2021)

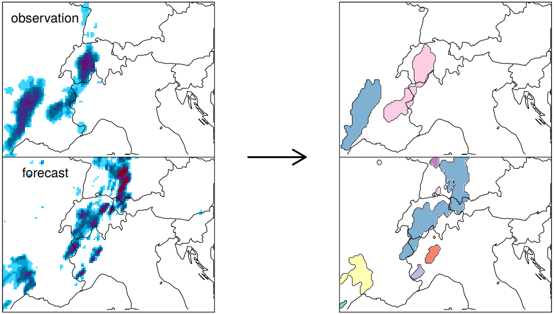
Ebert & McBride (2000)
Davis et al. (2006)
Wernli et al. (2008)
Radanovics et al. (2018)

Neighbourhood methods



Idea: Threshold fields (rain yes/no), apply smoothing, then compare grid-point wise. Repeat for different thresholds and smoothings.

Feature-based methods

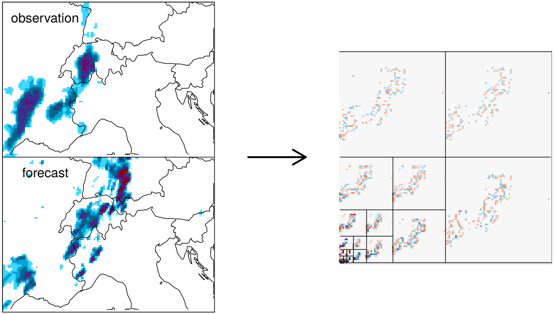


Idea: Decompose fields into discrete objects, then

(a) compare statistics of object properties

(b) match forecast and observed objects, compare directly

Scale-separation methods

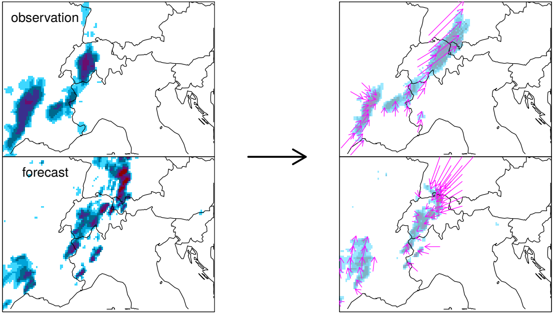


Idea: Decompose fields into components on different *scales*, then

(a) compare variance distribution across scales

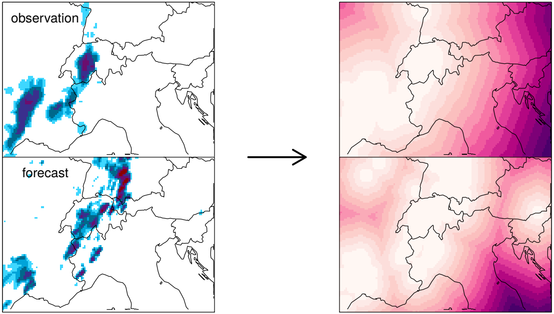
(b) compute grid-point wise errors on each scale

Field Deformation



Idea: Search for a vector field that transforms one field into the other, then measure the magnitude of the transformation.

Binary distance measures



Idea: Compute the distance to the nearest non-zero pixel (*distance map*) in each image, then compare.

Referenced Literature I

[1] Jeffrey L. Anderson.
A method for producing and evaluating probabilistic forecasts from ensemble model integrations.
Journal of Climate, 9:1518–1530, 1996.

[2] Louis Besson.
Essai de prévision méthodique du temps.
Annales de L'Observatiore Municipal, Ville de Paris, VI:473–495, 1905.

[3] Jochen Bröcker.
Reliability, sufficiency, and the decomposition of proper scores.
Quarterly Journal of the Royal Meteorological Society, 135(643):1512 – 1519, 2009.

[4] Jochen Bröcker.
Probability forecasts.
In Jolliffe and Stephenson [12], chapter 8, pages 119–139.

Referenced Literature II

[5] Jochen Bröcker.
Estimating reliability and resolution of probability forecasts through decomposition of the empirical score.
Climate Dynamics, 39(3-4):655–667, 2012.

[6] Jochen Bröcker.
Resolution and discrimination–two sides of the same coin.
Quarterly Journal of the Royal Meteorological Society, 141(689):1277–1282, 2015.

[7] Jochen Bröcker.
Testing the reliability of forecasting systems.
Journal of Applied Statistics, (accepted), 2021.

Referenced Literature III

[8] Jochen Bröcker and Zied Ben Bouallègue.
Stratified rank histograms for ensemble forecast verification under serial dependence.
Quarterly Journal of the Royal Meteorological Society, 146(729):1976–1990, 2020.

[Dawid(1984)] A. Philip Dawid.
Statistical theory. The prequential approach.
J. Roy. Statist. Soc. Ser. A, 147(2):278–292, 1984.

[DWD] DWD.
Surface temperature data from DWD weather stations.
Deutscher Wetterdienst, 2020.
URL `https://opendata.dwd.de/climate_environment/CDC/observations_germany/climate/hourly/air_temperature/historical`.

Referenced Literature IV

[ECMWF] ECMWF.
ECMWF operational archive.
European Centre for Medium Range Weather Forecasts, 2020.
URL https://www.ecmwf.int/en/forecasts/dataset/operational-archive.

[9] Raffaella Giacomini and Halbert White.
Tests of conditional predictive ability.
Econometrica, 74(6):1545–1578, 2006.

[Gneiting et al.(2007)Gneiting, Balabdaoui, and Raftery] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery.
Probabilistic forecasts, calibration and sharpness.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(2):243–268, 2007.

Referenced Literature V

[10] Thomas M. Hamill.
Interpretation of rank histograms for verifying ensemble forecasts.
Monthly Weather Review, 129(3):550–560, 2001.

[11] Thomas M. Hamill and Stephen J. Colucci.
Verification of Eta–RSM short range ensemble forecasts.
Monthly Weather Review, 125:1312–1327, 1997.

[12] Ian T. Jolliffe and David B. Stephenson, editors.
Forecast Verification; A practicioner's Guide in Athmospheric Science.
John Wiley & Sons, Ltd., Chichester, second edition, 2012.

[13] Natalia Nolde and Johanna F. Ziegel.
Elicitability and backtesting: perspectives for banking regulation.
Ann. Appl. Stat., 11(4):1833–1874, 2017.

Referenced Literature VI

[14] Olivier Talagrand, R. Vautard, and B. Strauss.
Evaluation of probabilistic prediction systems.
In *Workshop on Predictability*, pages 1–25. ECMWF, 1997.

[15] Andreas P. Weigel.
Verification of ensemble forecasts.
In Jolliffe and Stephenson [12], chapter 9, pages 141–166.

Further literature on scoring rules and reliability I

[1] Frédéric Atger.
Estimation of the reliability of ensemble based probabilistic forecasts.
Quaterly Journal of the Royal Meteorological Society, 130: 627–646, 2004.

[2] Herman J Bierens.
A consistent conditional moment test of functional form.
Econometrica: Journal of the Econometric Society, pages 1443–1458, 1990.

[3] Jochen Bröcker and Holger Kantz.
The concept of exchangeability in ensemble forecasting.
Nonlinear Processes in Geophysics, 18(1):1–5, 2011.
doi: 10.5194/npg-18-1-2011.

Further literature on scoring rules and reliability II

[4] Jochen Bröcker and Leonard A. Smith.
Increasing the reliability of reliability diagrams.
Weather and Forecasting, 22(3):651–661, June 2007.

[5] Robert M De Jong.
The Bierens test under data dependence.
Journal of Econometrics, 72(1-2):1–32, 1996.

[6] Francis X. Diebold and Jose A. Lopez.
Forecast evaluation and combination.
In *Statistical methods in finance*, volume 14 of *Handbook of Statist.*, pages 241–268. North-Holland, Amsterdam, 1996.
doi: 10.1016/S0169-7161(96)14010-4.

[7] Timo Dimitriadis, Tilmann Gneiting, and Alexander I Jordan.
Stable reliability diagrams for probabilistic classifiers.
Proceedings of the National Academy of Sciences, 118(8):
e2016191118, 2021.

Further literature on scoring rules and reliability III

[8] Timo Dimitriadis, Lutz Duembgen, Alexander Henzi, Marius Puke, and Johanna Ziegel.
Honest calibration assessment for binary outcome predictions.
arXiv preprint arXiv:2203.04065, 2022.

[9] Robert F. Engle and Simone Manganelli.
CAViaR: conditional autoregressive value at risk by regression quantiles.
J. Bus. Econom. Statist., 22(4):367–381, 2004.
ISSN 0735-0015.
doi: 10.1198/073500104000000370.

[10] Tobias Fissler and Johanna F Ziegel.
Higher order elicitability and osband's principle.
The Annals of Statistics, 44(4):1680–1707, 2016.

Further literature on scoring rules and reliability IV

[11] Wagner Piazza Gaglianone, Luiz Renato Lima, Oliver Linton, and Daniel R. Smith.
Evaluating value-at-risk models via quantile regression.
J. Bus. Econom. Statist., 29(1):150–160, 2011.
ISSN 0735-0015.
doi: 10.1198/jbes.2010.07318.

[12] Tilmann Gneiting.
Making and evaluating point forecasts.
Journal of the American Statistical Association, 106(494): 746–762, 2011.
doi: 10.1198/jasa.2011.r10138.

Further literature on scoring rules and reliability V

[13] Tilmann Gneiting and Johannes Resin.
Regression diagnostics meets forecast evaluation: Conditional calibration, reliability diagrams, and coefficient of determination.
arXiv, 2021.
doi: 10.48550/ARXIV.2108.03210.

[14] Jacob A. Mincer and Victor Zarnowitz.
The evaluation of economic forecasts.
In Jacob A. Mincer, editor, *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, pages 3–46. National Bureau of Economic Research, 1969.
ISBN 0-870-14202-X.

Further literature on scoring rules and reliability VI

[15] Ingo Steinwart, Chloé Pasin, Robert Williamson, and Siyu Zhang.
Elicitation and identification of properties.
In M.F. Balcan, V. Feldman, and C. Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 482–526, Barcelona, Spain, 13–15 Jun 2014. PMLR.

[16] Daniel S. Wilks.
Statistical Methods in the Athmospheric Sciences, volume 59 of *International Geophysics Series*.
Academic Press, first edition, 1995.

Further literature on spatial verification I

General overview:

Eric Gilleland, David Ahijevych, et al. "Intercomparison of Spatial Forecast Verification Methods". In: *Weather and Forecasting* (Oct. 2009). DOI: 10.1175/2009WAF2222269.1

Manfred Dorninger et al. "The Setup of the MesoVICT Project". In: *Bulletin of the American Meteorological Society* (Sept. 2018). DOI: 10.1175/BAMS-D-17-0164.1

Sebastian Buschow. "Spatial Verification with Wavelets". PhD thesis. Rheinische Friedrich-Wilhelms-Universität Bonn, Apr. 2022. URL: https://hdl.handle.net/20.500.11811/9723

Further literature on spatial verification II

Neighbourhood methods:

Nigel M Roberts and Humphrey W Lean. "Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events". In: *Monthly Weather Review* (2008). DOI: https://doi.org/10.1175/2007MWR2123.1

Elizabeth E Ebert. "Fuzzy verification of high-resolution gridded forecasts: A review and proposed framework". In: *Meteorological Applications* (2008). DOI: https://doi.org/10.1002/met.25

Gregor Skok and Veronika Hladnik. "Verification of Gridded Wind Forecasts in Complex Alpine Terrain: A New Wind Verification Methodology Based on the Neighborhood Approach". In: *Monthly Weather Review* (Jan. 2018). DOI: 10.1175/MWR-D-16-0471.1

Further literature on spatial verification III

Feature-based methods:

EE Ebert and JL McBride. "Verification of precipitation in weather systems: Determination of systematic errors". In: *Journal of hydrology* (2000). DOI: https://doi.org/10.1016/S0022-1694(00)00343-7

Heini Wernli et al. "SAL—A Novel Quality Measure for the Verification of Quantitative Precipitation Forecasts". In: *Monthly Weather Review* (2008). DOI: 10.1175/2008MWR2415.1

Christopher A. Davis et al. "The Method for Object-Based Diagnostic Evaluation (MODE) Applied to Numerical Forecasts from the 2005 NSSL/SPC Spring Program". In: *Weather and Forecasting* (2009). DOI: 10.1175/2009WAF2222241.1

Further literature on spatial verification IV

Scale-separation methods:

B Casati et al. "A new intensity-scale approach for the verification of spatial precipitation forecasts". In: *Meteorological Applications* (2004). DOI: https://doi.org/10.1017/S1350482704001239

Jun-Ichi Yano and Bogumil Jakubiak. "Wavelet-based verification of the quantitative precipitation forecast". In: *Dynamics of Atmospheres and Oceans* (2016). DOI: https://doi.org/10.1016/j.dynatmoce.2016.02.001

Michael Weniger et al. "Spatial verification using wavelet transforms: A review". In: *Qarterly Journal of the Royal Meteorological Society* (2017). DOI: https://doi.org/10.1002/qj.2881

Further literature on spatial verification V

More scale-separation methods:

Sebastian Buschow and Petra Friederichs. "SAD: Verifying the scale, anisotropy and direction of precipitation forecasts". In: *Quarterly Journal of the Royal Meteorological Society* (2021). DOI: https://doi.org/10.1002/qj.3964

Sebastian Buschow. "Measuring displacement errors with complex wavelets". In: *Weather and Forecasting* (2022). DOI: https://doi.org/10.1175/WAF-D-21-0180.1

Barbara Casati et al. "Scale-separation diagnostics and the Symmetric Bounded Efficiency for the inter-comparison of precipitation reanalyses". In: *International Journal of Climatology* (2023). DOI: https://doi.org/10.1002/joc.7975

Further literature on spatial verification VI

Field Deformation:

Christian Keil and George C Craig. "A displacement and amplitude score employing an optical flow technique". In: *Weather and Forecasting* (2009). DOI: https://doi.org/10.1175/2009WAF2222247.1

Alban Farchi et al. "Using the Wasserstein distance to compare fields of pollutants: Application to the radionuclide atmospheric dispersion of the Fukushima-Daiichi accident". In: *Tellus B: Chemical and Physical Meteorology* (2016). DOI: https://doi.org/10.3402/tellusb.v68.31682

Fan Han and Istvan Szunyogh. "A Technique for the Verification of Precipitation Forecasts and Its Application to a Problem of Predictability". In: *Monthly Weather Review* (2018). DOI: https://doi.org/10.1175/MWR-D-17-0040.1

Further literature on spatial verification VII

<u>Distance measures</u>:

Eric Gilleland. "Spatial forecast verification: Baddeley's delta metric applied to the ICP test cases". In: *Weather and Forecasting* (2011). DOI: https://doi.org/10.1175/WAF-D-10-05061.1

Eric Gilleland, Gregor Skok, et al. "A Novel Set of Geometric Verification Test Fields with Application to Distance Measures". In: *Monthly Weather Review* (2020). DOI: https://doi.org/10.1175/MWR-D-19-0256.1