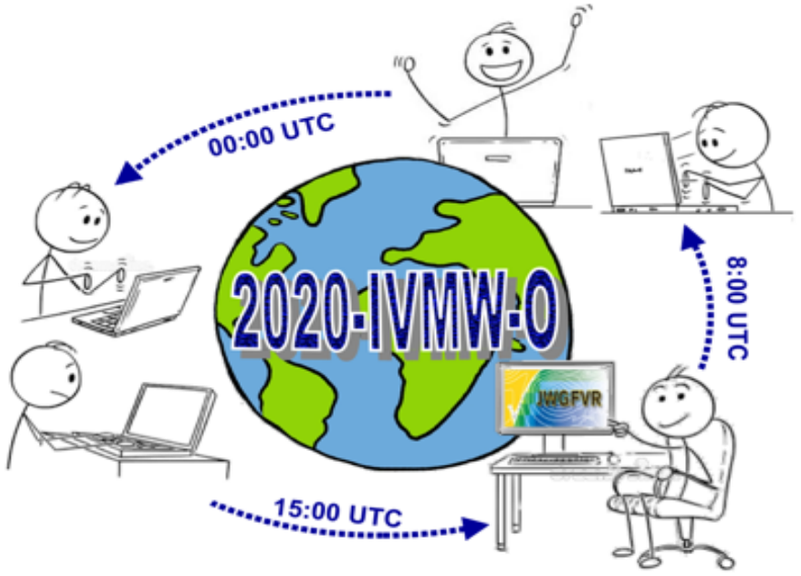
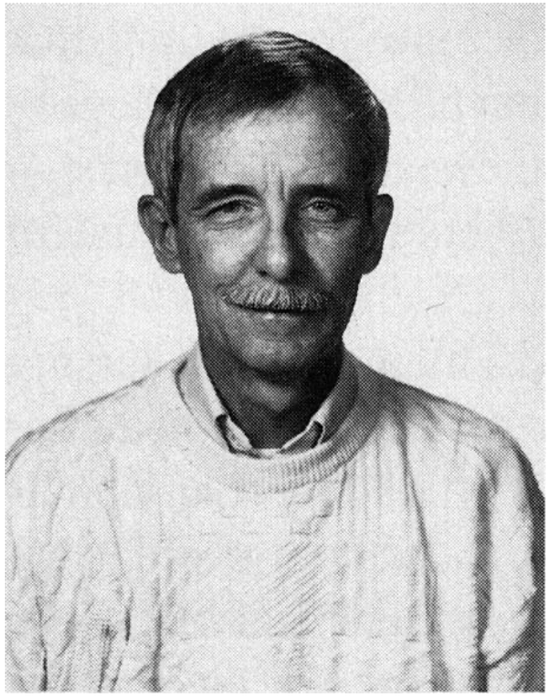
A 20-year Journey of Forecast Verification Research

Barbara Casati (MRD/ECCC, Canada)
Caio Coelho (CPTEC/INPE, Brazil)
Manfred Dorninger (University of Vienna, Austria)
Beth Ebert (Bureau of Meteorology, Australia)
Eric Gilleland (NCAR/RAL, USA)
Chiara Marsigli (DWD, Germany)
Marion Mittermaier (MetOffice, UK)



The Funds: Murphy's Law



Verification: process which compares forecast with obs

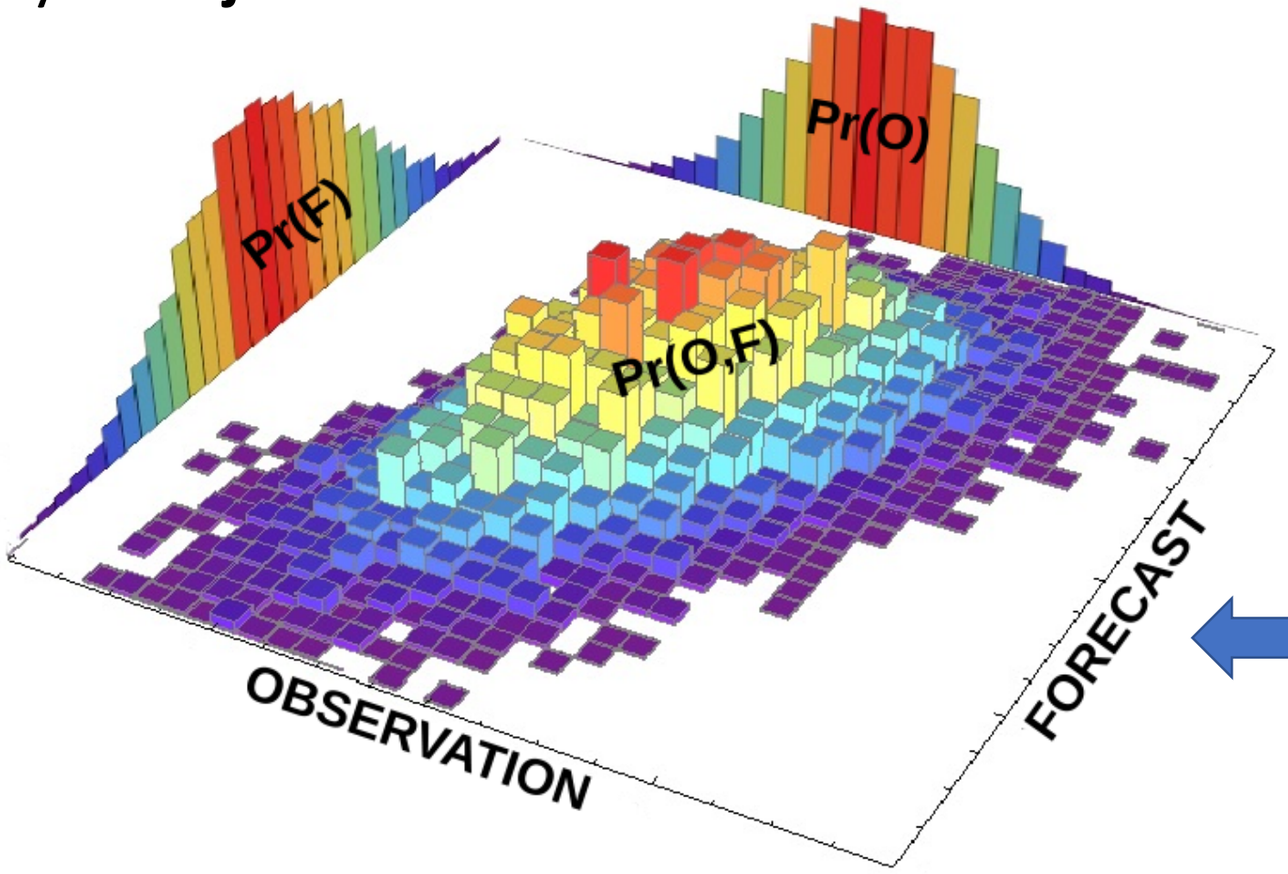
Why verify?
- **Administrative**: monitor forecast, compare different model versions
- **Scientific**: identify model weaknesses, improve model physics
- **Economic**: guide decisions of forecast end-users

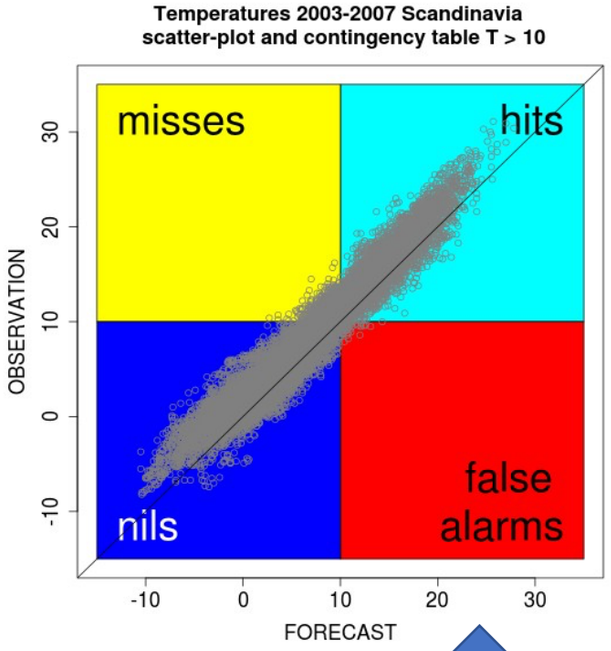
Verification measures the **quality** of the forecast (as opposed to **value** and **evaluation**). Several **attributes** of the **quality**: **bias, accuracy, association, skill ...** Scores provide a summary measure of these attributes (e.g. ME describes the bias, MSE measures the accuracy, corr coefficient measure linear association, ...).

Murphy (1993) *What is a good forecast? An essay on the nature of goodness in weather forecasting*. **Murphy (1991)** *Forecast verification: Its complexity and dimensionality;* **Murphy and Winkler (1987)**: *A general framework for forecast verification.*

The Funds: joint distribution

Verification is intrinsically a statistical problem: the forecast-observation relationship is fully described by their **joint distribution**





Predicted \ Observed	O	O^c	
F	$Pr(O,F)$	$Pr(O^c,F)$	$Pr(F)$
F^c	$Pr(O,F^c)$	$Pr(O^c,F^c)$	$Pr(F^c)$
	$Pr(O)$	$Pr(O^c)$	1

A 20-year journey of verification research

st

- Wilks (1995), (2006), (20
- Jolliffe and Stephenson (

Verification Workshops

- 30 July – 1 Aug **2002, Boulder**: “Making Verification More Meaningful” (Barb Brown).
- 15-17 Sept **2004, Montreal**: 2nd International Verification Workshop (Laurie Wilson)
- 31 Jan – 2 Feb **2007, Reading**: 3rd International Verification Workshop & Tutorials (Anna Ghelli)
  https://www.ecmwf.int/en/learning/workshops-and-seminars/past-workshops/2007-international-verification-methods
  Ebert and Ghelli (2008) ed. Met Apps Special Issue;
  Casati et al (2008) review article
- 4 -10 June **2009, Helsinki**: 4th International Verification Workshop & Tutorials (Pertti Nurmi)
  https://space.fmi.fi/Verification2009/
- 1-7 Dec **2011, Melbourne**: 5th International Verification Workshop & Tutorials (Beth Ebert)
  Ebert et al (2013) review article
- 13-19 March **2014, New Delhi**: 6th International Verification Workshop & Tutorials (Raghu Ashrit)
- 3-11 May **2017, Berlin**: 7th International Verification Workshop & Tutorials (Martin Goeber)
  https://www.7thverificationworkshop.de
  Dorninger et al (2018) ed. Met Zet special issue;
  Dorninger et al (2020) ed. Met Apps special issue.
- 9-20 November **2020, Online** International Verification Method Workshop (Barbara Casati)

Why do we want to verify?

E.Ebert (2008): "**Forecast is like an experiment** – given a set of conditions one makes an hypothesis that a certain outcome will occur. **The experiment is not completed until its outcome is determined!** This is the act of verification: determining whether the forecast was successful"

=> **Verification** has become an **integral part of the forecasting process**.

Murphy (1993) verification purposes:

- **Administrative**: monitor forecast, compare different model versions
- **Scientific**: identify model weaknesses, improve model physics
- **Economic**: guide decisions of forecast end-users

=> Jolliffe and Stephenson (2003): **The scope of verification is informative**: enhance understanding on the forecast capabilities, gain knowledge on the forecast performance.

=> Verification: ultimate goal is to **use the gained information for improvements**!

- Upstream = Enhance understanding for guiding new NWP developments/improvements.
- Downstream = gain knowledge for an informed and better use of forecast products.

Which score? Which method?

1. Graphical summary (scatter-plot, qq-plot, …)
2. Continuous scores (MSE,correlation, …)
3. Categorical scores (FBI,HSS,PC, …).
4. Probabilistic scores (Brier,CRPS, …).
5. Extremes (EDI,SEDI)
6. Spatial methods: Scale-separation; Neighbourhood; Field-deformation; Feature-based; Distance metrics
7. SEEPS,Generalized Discrimination Score, Minimum Spanning Tree, …

There is no single technique which fully describes the complex observation-forecast relationship, no single score can provide a complete picture: explore!!!
Almost always we need a portfolio of statistics … the more the merrier :o)

Key Q1: what do we wish to know from our verification ?
(end-user / verification purpose / questions addressed / attributes of interest)
Key Q2: What are the (statistical) characteristics
of the variable and forecast verified?
Key Q3: What are the available observations?

Key factors to define the verification strategy

Purpose of the verification

Variable of interest

Forecast characteristics

Available observations

Key factors to define the verification strategy

Purpose of the verification

Variable of interest

Forecast characteristics

Available observations

- Summary performance measures for monitoring NWP improvements, comparing competing forecasting systems => **Confidence Intervals** (Mason 2008, Jolliffe 2007)
- Selecting, clustering, blending multi-model ensembles / multiple NWP products.
- **Physically meaningful diagnostics for model developers**.
- The quest for model systematic errors: post-processing, downscaling and bias-correction of NWP model outputs.
- Evaluation of the added value of enhanced resolution models => spatial methods
- Robust and informative metrics for extremes and high impact weather.
- **User-oriented** indices to help planning response to weather conditions: impact and adaptation strategy.

Key factors to define the verification strategy

Purpose of the verification

Variable of interest

Forecast characteristics

Available observations

- Summary performance measures for monitoring NWP improvements, comparing competing forecasting sys
- Selecting, clustering, blen
- **Physically meaningful d**
- The quest for model syste correction of NWP model
- Evaluation of the added v
- Robust and informative m
- **User-oriented** indices to adaptation strategy.

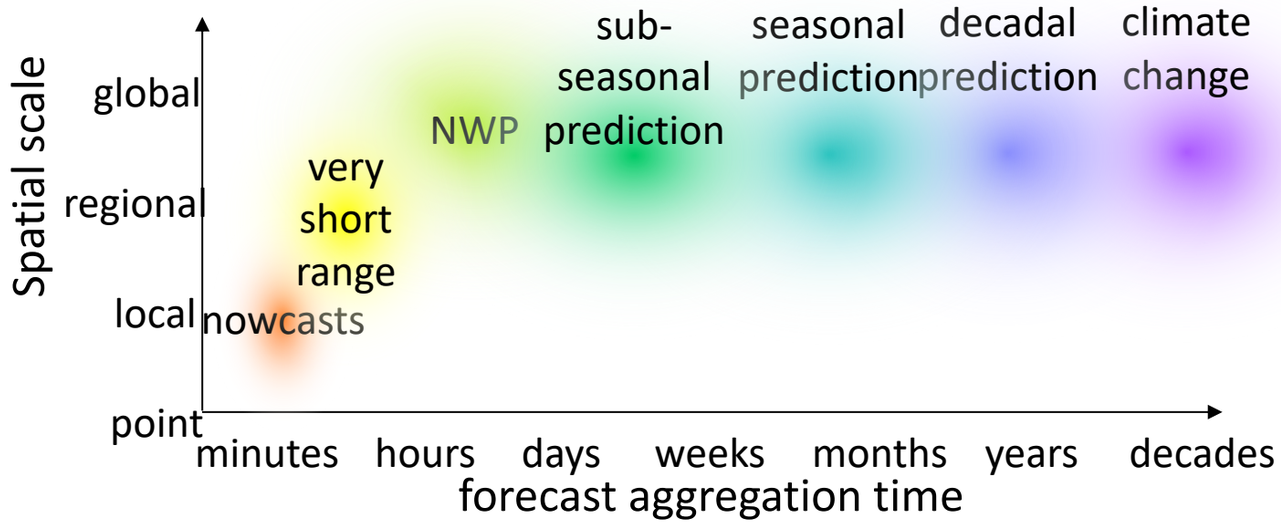
Advanced forecast verification end-users: hydrology, agriculture, energy sector. Transport: road conditions, air-transport, navigation (e.g. in sea-ice infested waters). Urban modeling (e.g. increasing heat islands with climate change) to enable better planning of urban infrastructure.

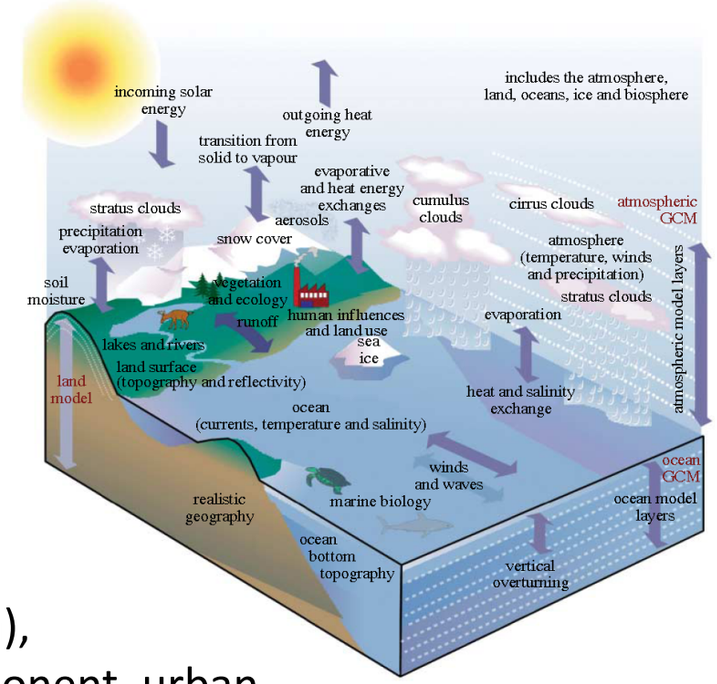
The modellers and forecasters: a very special class of users, since they are upstream in the science-to-services value chain (as opposed to the previous users, which are downstream in the chain).

Key factors to define the verification strategy

Purpose of the verification	Variable of interest	Forecast characteristics	Available observations

Increasing complexity and dimensionality of NWP: ensembles and higher resolution models; spatial and time scales: from short-medium to long range, monthly to seasonal, decadal and climate **=> S2S,S2D, Climate**





From NWP to Earth System Models: surface component, land & vegetation, cryosphere (snow and glaciers), sea-ice and ocean (tides and waves, coastal erosion), hydrology (run-off, rivers), troposphere to stratosphere, chemical and aerosol component, urban …

Key factors to define the verification strategy

Purpose of the verification

Variable of interest

Forecast characteristics

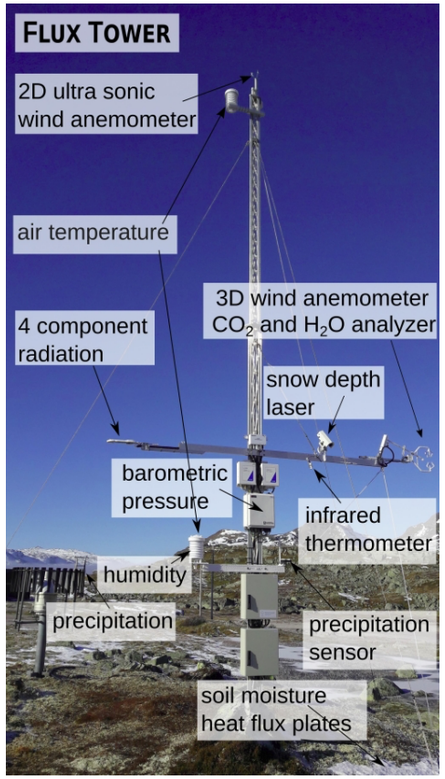
Available observations

<u>End-users</u>: transport: fog and visibility, icing, sea-ice pressure; agriculture: cold/warm degree days; infrastructure: erosion; …

<u>Scientific communities</u>: sea-ice and ocean (e.g. salinity, wave height); surface: soil moisture, snow depth; hydrology: run-off and spring flood;

<u>Multiple sensors and observatories</u>: variable interactions, physical processes.

Enhanced station networks
Radars and satellite products



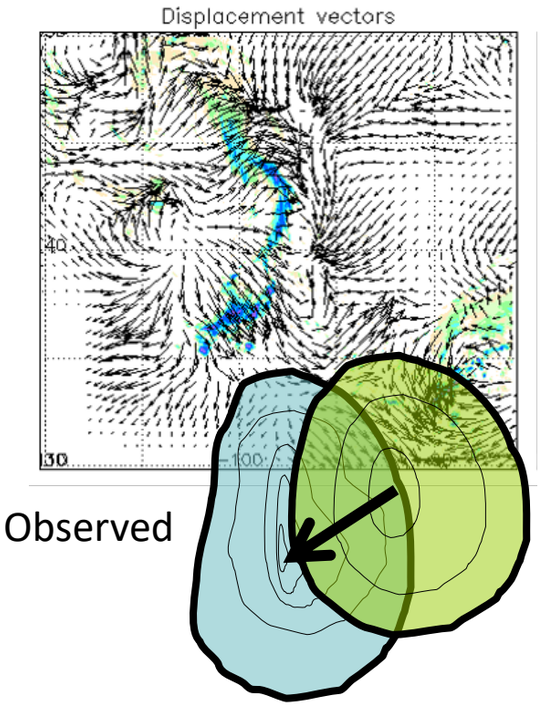
New sensors: fluxes, clouds and precip microphysics, …

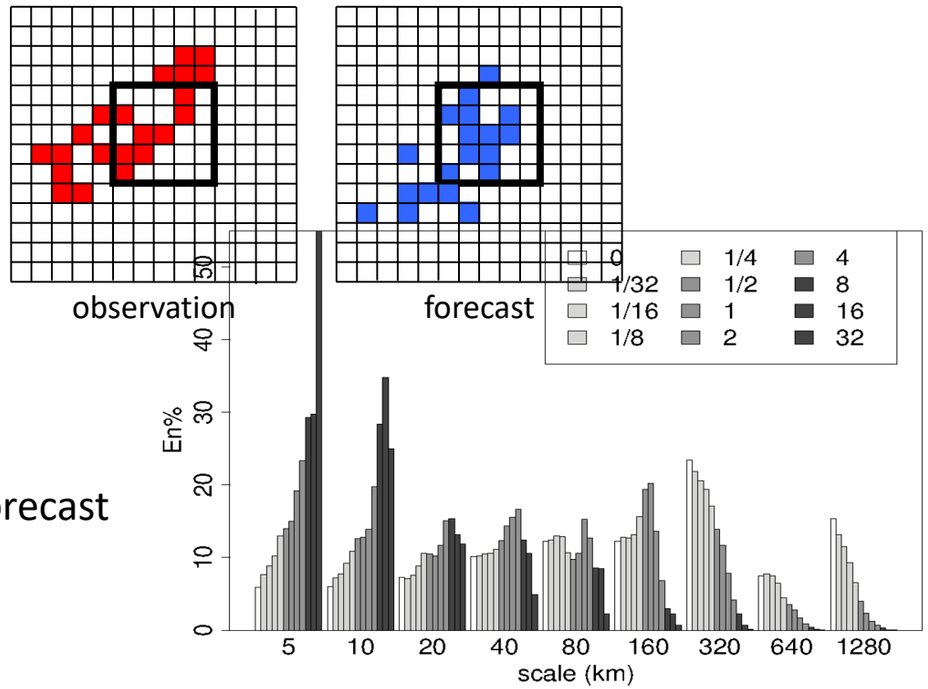
New observations: crowdsourcing and citizen science (public) initiatives

Spatial verification methods

- Account for **coherent spatial structure** and the presence of **features**
- Aim to provide information on **error in physical terms (meaningful verification): e.g.** assess **scale structure** and **displacement error** (separately from **intensity error**)
- Account for **small time-space uncertainties** (avoid **double-penalty** issue)

The origins: Evaluation of the added value of enhanced resolution in QPF: CRA (Ebert and McBride, 2000; Grams et al 2006); IS (Casati et al 2004); FSS (Roberts and Lean 2008), MODE (Brown et al, 2004; Davis et al, 2006); SAL (Wernli, 2006); DAS (Keil and Craig, 2008) ...

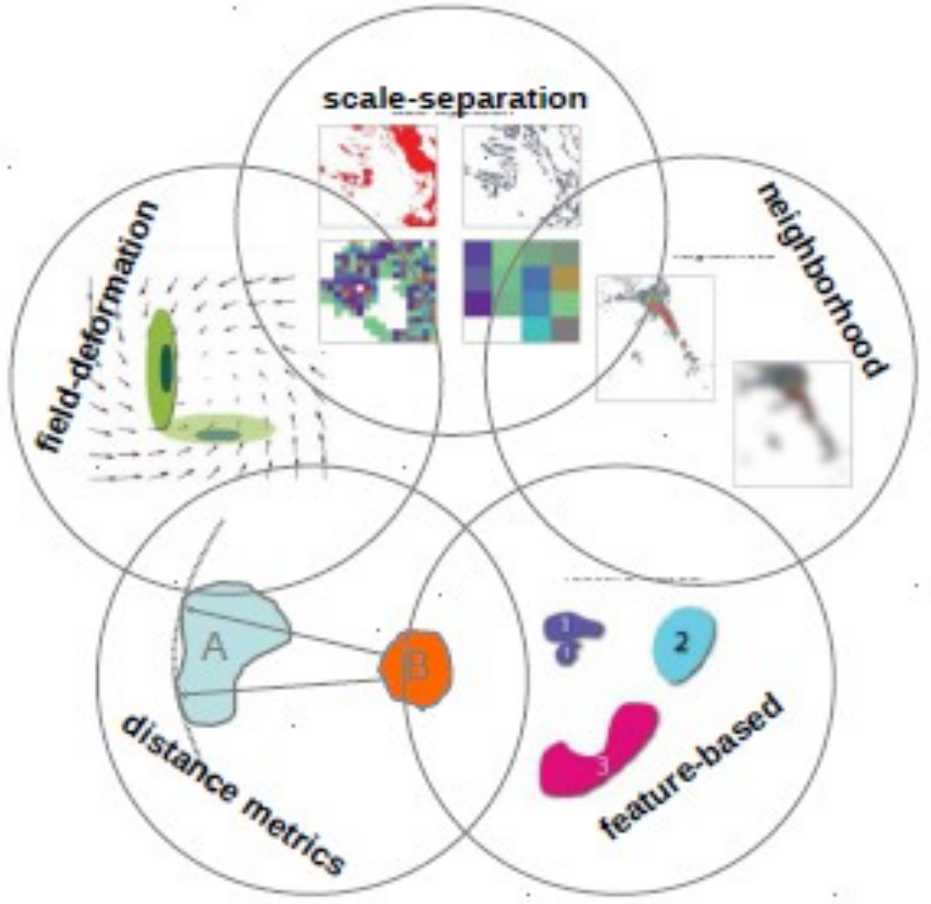




Spatial verification methods Inter-Comparisons

- Spatial Verification Inter-Comparison Project (ICP)
  Gilleland et al (2010), BAMS
- Mesoscale Verification Intercomparison in complex Terrain (MesoVICT)
  Dorninger et al (2018), BAMS

Analyse the behaviour and classify the spatial methods: provide guidance

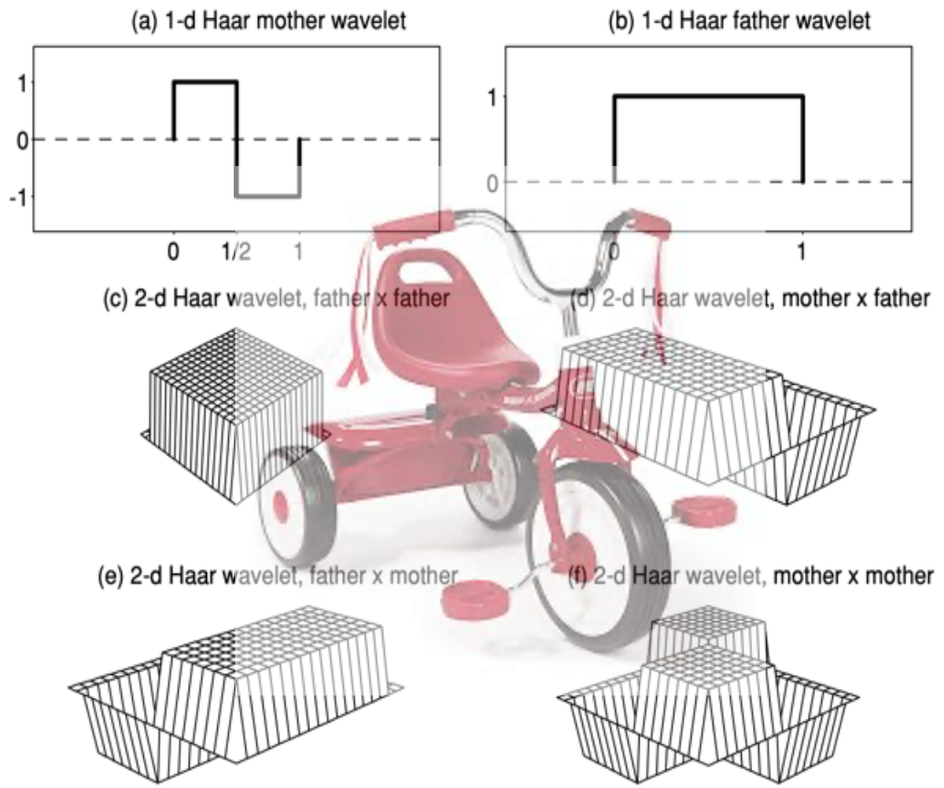


http://www.ral.ucar.edu/projects/icp
Includes an impressive list of references, more than 200 peer-review articles.

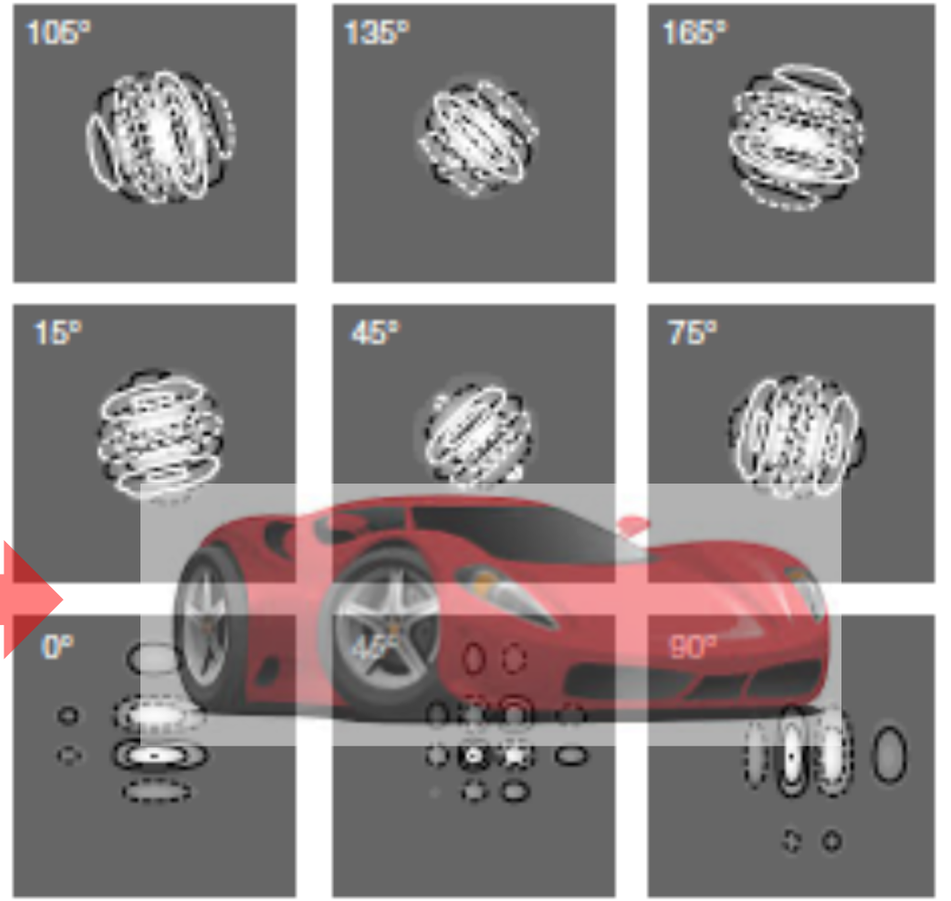
Open source community verification tools:
R packages, MET and METplus

The evolution of spatial verification methods

One example,
from the scale separation techniques:



From Casati et al (2004) use Discrete Wavelet Transform with Haar wavelets -> “bloky” features



From Buschow and Friederichs (2020) QJRMS, DTCWT versus DWT: characterize orientation angle and anisotropy, further than the spatial scale.

Sea-ice verification: distance and areal metrics

Several products from **satellites => gridded observations!**
Several attributes: sea ice concentration, thickness, pressure, …

Sea ice edge:
Melsom et al (2019), Zampieri (2018), inter-comparisons
Goessling and Jung (2018) spatial probabitlity score **(SPS)**
Goessling et al (2016) integrated ice edge error **(IIEE)**
Dukhovskoy et al (2015), mean distances and ModHausDist
Heinrichs et al (2006), Frechet distance

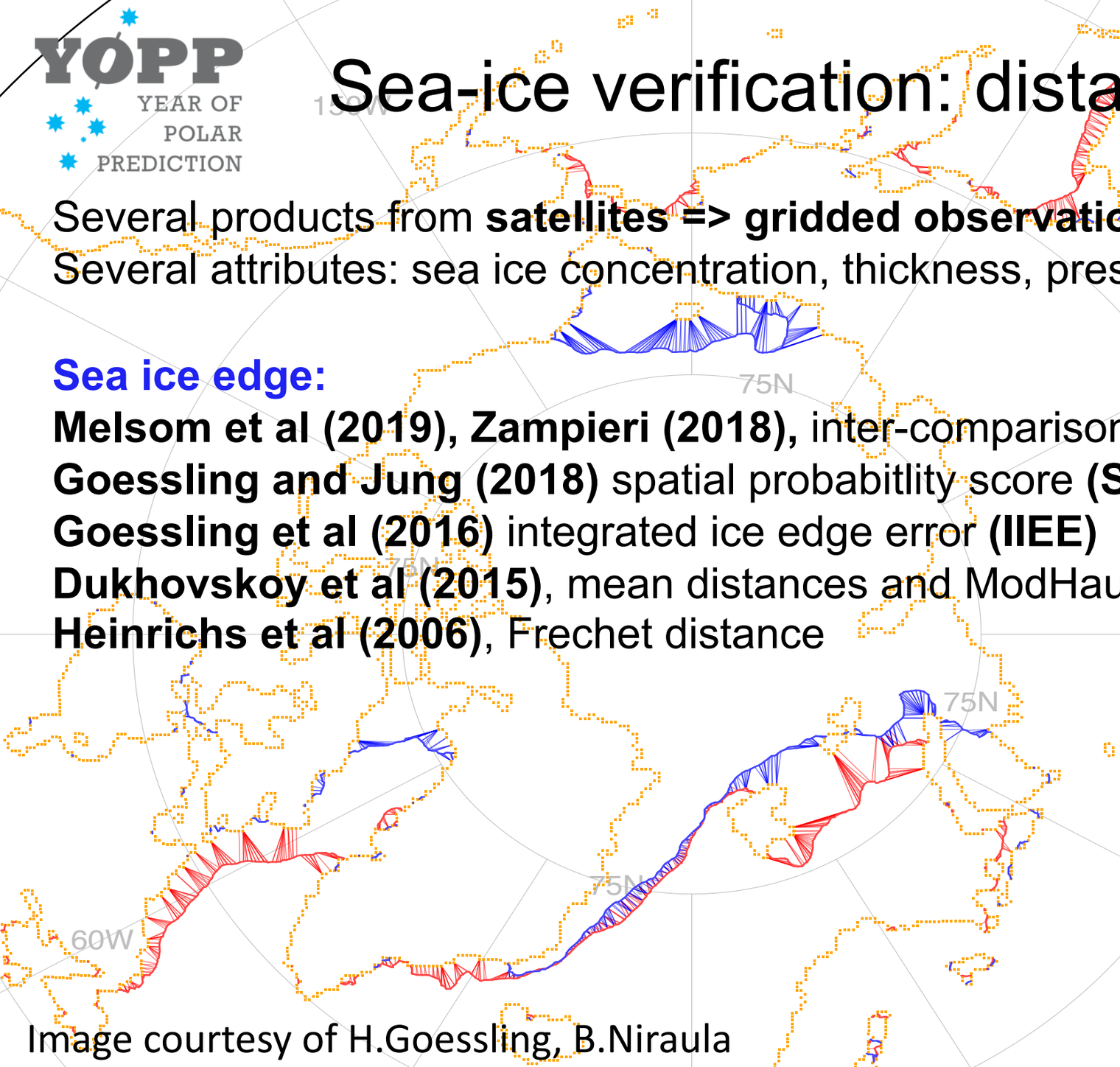
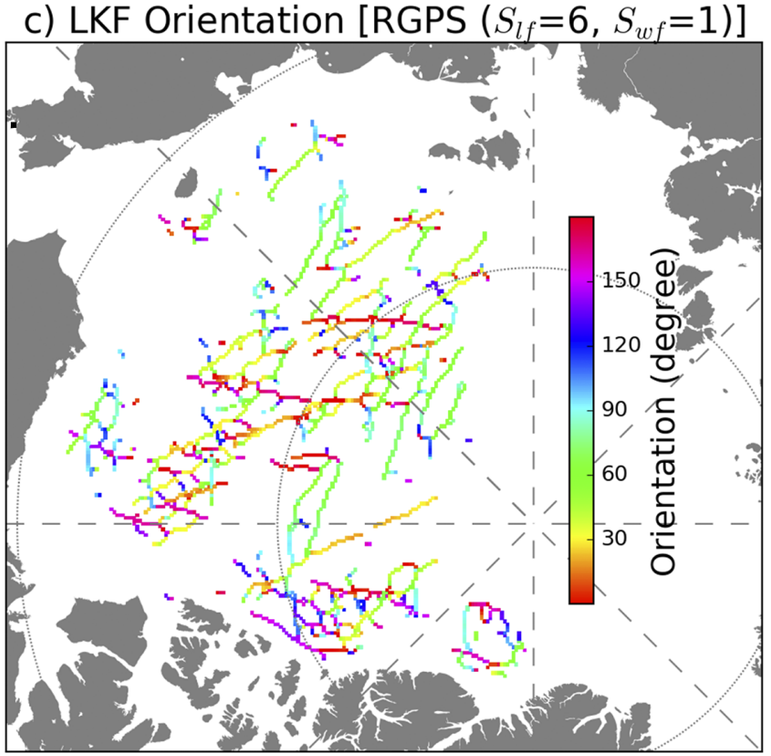


Image courtesy of H.Goessling, B.Niraula



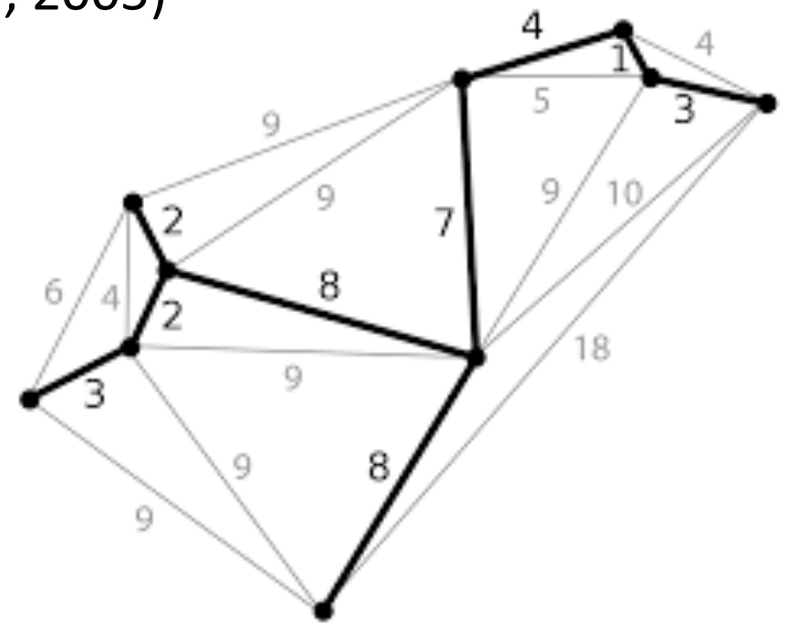
Linear Kinematik Features:
Mohammadi-Aragh et al (2020)
multiscale directional analysis
Linow and Dierking (2017)
Object-based detection of LKF

Metaverification

Analyse and challenge existing verification scores, to better understand strengths and limits of each verification measure. Often leads to the development of new scores.

Recent progress gravitate mostly around ensembles and prob forecasts verification:

- Brier and CRPS -> binning and sample size (Ferro 2007, Ferro et al 2008, Stephenson et al 2008)
- Rank histogram -> binning and sample size (Candille and Talagrand, 2005)
- property -> Broker and Smith (2007), Gneiting and Raftery (2007)
- hedging: property equality consistency -> Jolliffe (2008)
- ETS is not equitable! -> Hogan et al (2010)
- Rodwell et al (2010) introduce Stable Equitable Error in Probability Space (SEEPS)



Ensembles: CRPS most accepted summary measure,
along with Reliability diagram and ROC curve (discrimination);

- Minimum Spanning Tree (Smith and Hansen, 2004; Wilks 2004);
- Discrimination Generalized Score (Weigel and Mason, 2011; Mason and Weigel, 2009).

Rodwell et al (2010), SEEPS

Introduction: clearly define the desired characteristics of the sought verification measure:

1. Single summary measure
2. Use station observations
3. Accounts for local climatology
4. Meaningful geographical aggregation
5. Deals with precipitation mix distribution (dry & precip values)
6. Small sensitivity to sample uncertainty
7. Discrimination
8. Refinement
9. Proper and equitable

Here the aim is to develop a new score that concisely quantifies NWP performance in the prediction of precipitation and steers development in the correct direction. The desirable attributes of such a score can be summarized as follows.

(a) Monitoring Progress.

- A single score should be sought that assesses forecast skill for dry weather and precipitation quantity.
- Verification against point observations is required in order to permit continuous monitoring of a system with resolution increasing with time, and to satisfy the typical user interested in a small geographic area.
- To detect performance changes, sensitivity to sampling uncertainty should be minimized, while maintaining the ability to differentiate between 'good' and 'bad' forecasts.
- For area and temporal averages to be meaningful, it should be possible to aggregate scores from different climate regions and different times of the year.

(b) Aiding decision-making.

- To facilitate the identification of model error, it should be possible to plot a map of scores for a single forecast.
- A score should encourage developments that permit a forecast system to predict the full range of possible outcomes.
- A better score should indicate a 'better forecast system'.

Rodwell et al (2010), SEEPS

Nicely explain the equitability constraint (Gandin and Murphy 1992, Gerrity 1992), build step by step score with desired properties.

1. Write the multi-categorical score as sum of joint probabilities $[p_{i,j}]$ weighted by elements of a scoring matrix $[s_{i,j}]$

$$Score = \sum_{i,j} s_{i,j} \cdot p_{i,j}$$

$p_{1,1}$	$p_{2,1}$
$p_{1,2}$	$p_{2,2}$

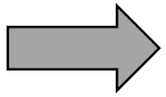
$\cdot$

$s_{1,1}$	$s_{2,1}$
$s_{1,2}$	$s_{2,2}$

2. Apply equitability constraints + symmetry

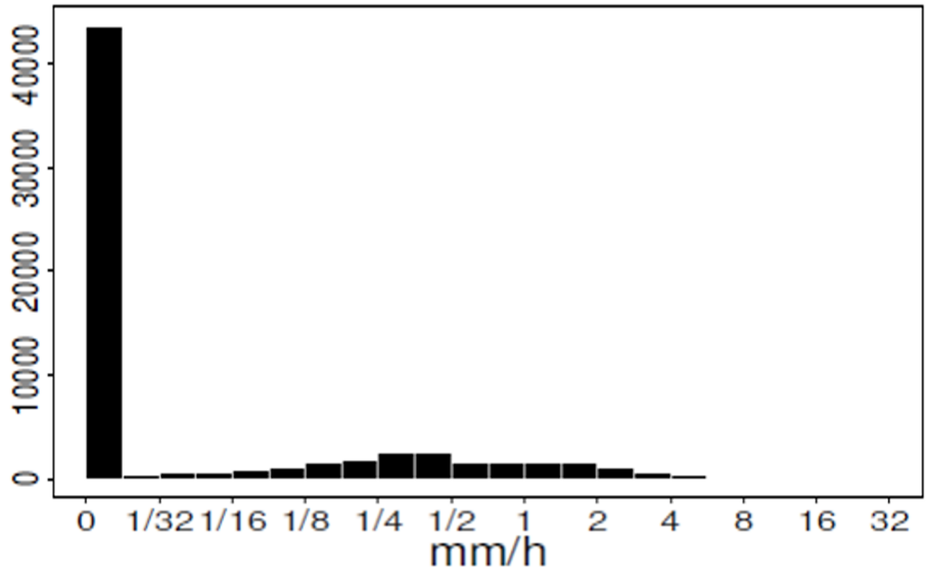
$$\begin{cases} s_{1,1}p_1 + s_{2,2}p_2 = 1 & \text{Perfect} \\ s_{1,1}p_1 + s_{2,1}p_2 = 0 & \text{Constant = 1} \\ s_{1,2}p_1 + s_{2,2}p_2 = 0 & \text{Constant = 2} \\ s_{2,1} = s_{1,2} & \text{symmetry} \end{cases}$$

3. Find scoring matrix for 2x2 equitable symmetric score



p_2/p_1	-1
-1	p_1/p_2

Rodwell et al (2010), SEEPS



Suitable for precipitation right skewed mix distribution (dry & precip values) => probability space (robust and resistant)

Accounts for local climatology => Enables a meaningful geographical aggregation, avoid false skill issue (Hamill and Juras, 2006)

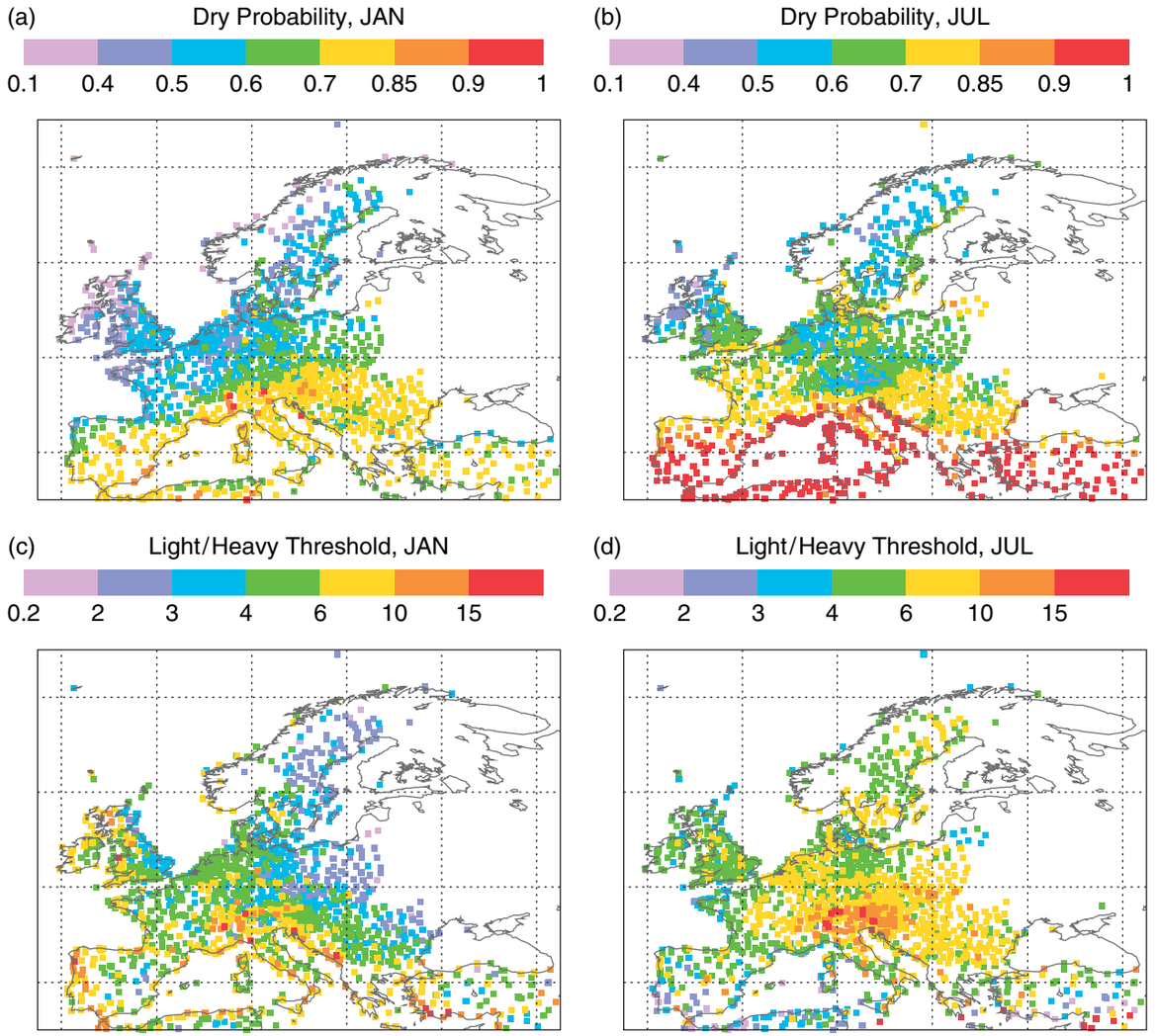


Figure 6. (a) Probability of a 'dry' day for January. (b) As (a) but for July. (c) Precipitation amount (in mm) marking the threshold between 'light' and 'heavy' precipitation for January. (d) As (c) but for July. By definition, 'light precipitation' occurs twice as often as 'heavy precipitation'. Results are based on 24 hour precipitation accumulations (1200 UTC–1200 UTC) from the 1980–2008 climatology.

False skill

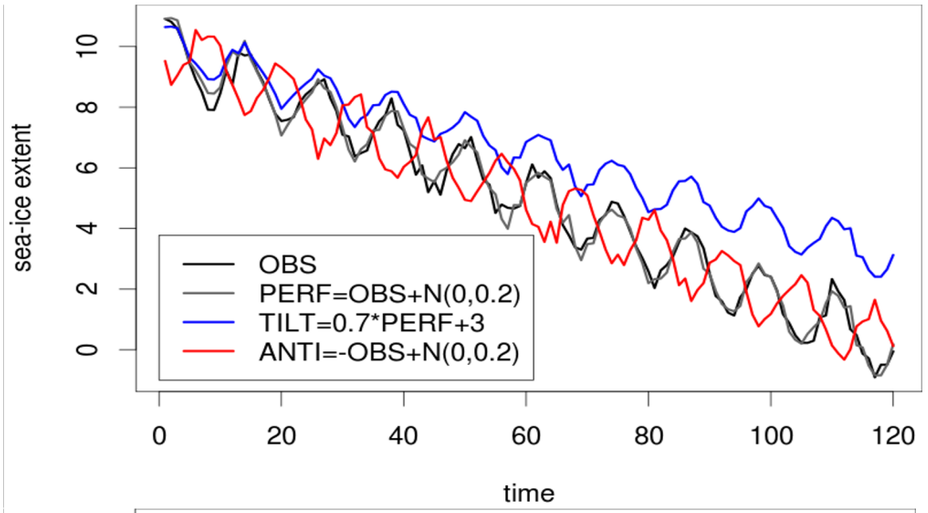
Hamill and Juras (2006): operational practice aggregate different locations. If climatologies of verified sample are different, part of the skill is due to reproducing the local climatologies.

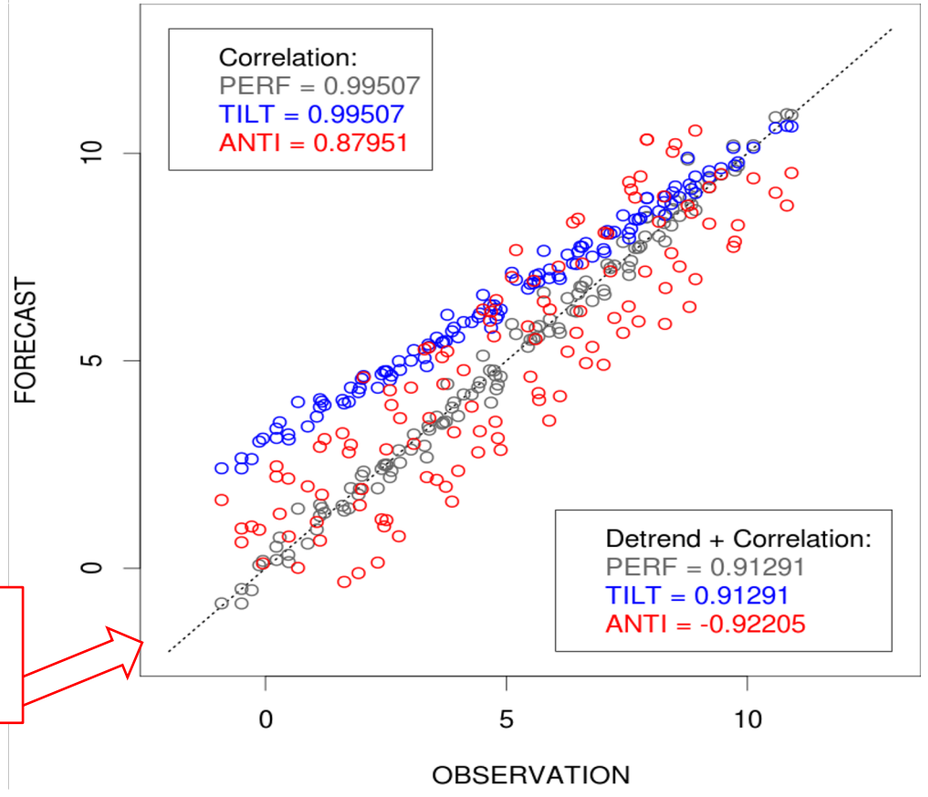
False (trivial) skill can also be scored for variables affected by climate trends

Ferro et al (2013) introduce a class of performance measures that are immune to spurious skill due to the presence of climate trends.

Correlation is not sensitive to bias, is not the regression line slope, is artificially inflated for variables with a (climate) trend

Sea-ice extent is characterized by annual cycle + decreasing trend





Challenging verification scores for extreme/rare events: from the Finley's affair (Murphy, 1996) to the extreme dependence indices debate.

Stephenson et al (2008): "The Extreme Dependency Score: a non-vanishing measure for forecasts of rare events"

Primo and Ghelli (2009), Ghelli and Primo (2009): the EDS is not proper and can be hedged by over-forecasting.

Hogan et al (2010): to address some of the shortcoming of the EDS they introduce the Symmertic EDS.

Ferro and Stephenson (2011): Revisit the properties of EDS and SEDS, introduce two new extremal dependence indices (EDI, SEDI).

from traditional categorical scores to bivariate EVT association measures

TABLE 6. Properties of five verification measures.

	ETS	EDS	SEDS	EDI	SEDI
Nondegenerate limit	×	✓	✓	✓	✓
Base-rate independent	×	×	×	✓	✓
Nontrivial to hedge	✓	×	✓	✓	✓
Regular	×	×	×	✓	✓
Fixed range [−1, 1]	×	×	×	✓	✓
Asymptotically equitable	✓	×	✓	✓	✓
Meaningful origin	✓	×	✓	✓	✓
Complement symmetric	✓	×	×	×	✓
Transpose symmetric	✓	×	✓	×	×

Asymptotic behaviour of the joint (bivariate) distribution

For rare/extreme events the base rate $\varepsilon \to 0$, then traditional scores degenerate to trivial non-informative limits (e.g. zero)

Scores defined by **logarithms**: the asymptotic behaviour depends on the rate of convergence to zero (α)

Predicted \ Observed	O	O^c	
F	ε^{α}	ε	ε
F^c	ε	$1-\varepsilon$	$1-\varepsilon$
	ε	$1-\varepsilon$	1

$$EDS = \frac{logp - logH}{logp + logH} \qquad SEDS = \frac{logq - logH}{logp + logH}$$

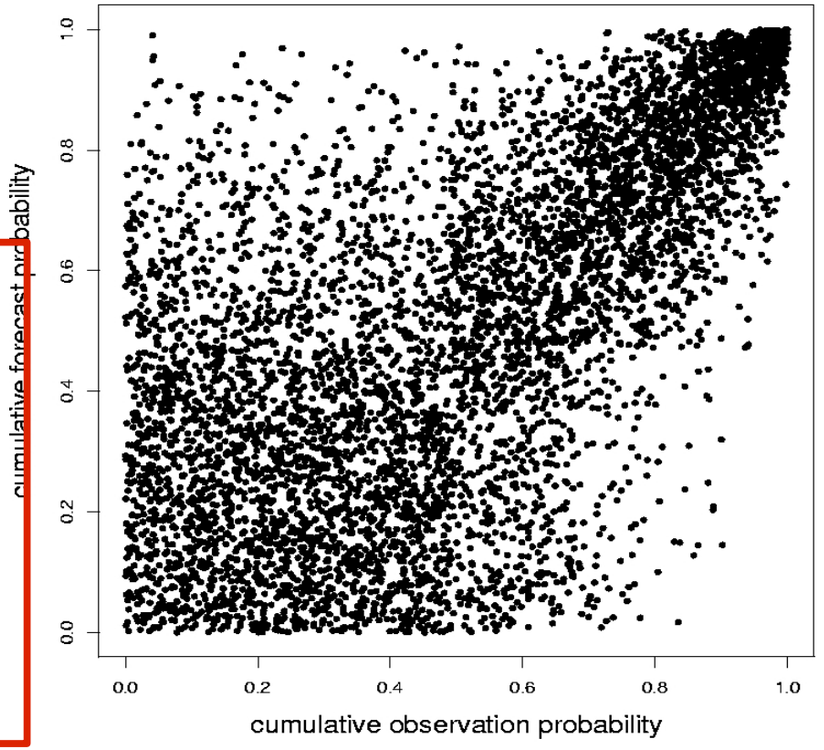
$$EDI = \frac{logF - logH}{logF + logH} \qquad SEDI = \frac{logF - logH - \log(1-F) + \log(1-H)}{logF + logH + \log(1-F) + \log(1-H)}$$

$$p = (a+c)/n \; q = (a+n)/n \; H = a/(a+c) \; F = b/(b+d)$$

Extreme Value Theory (EVT)

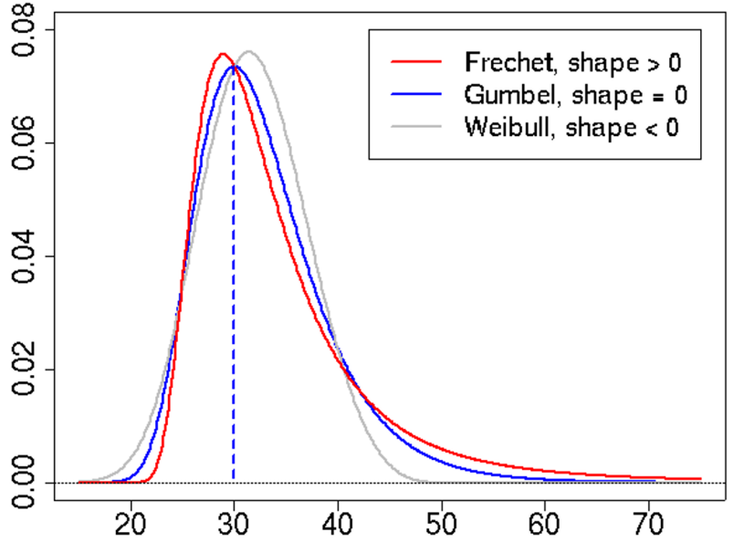
The EDS is an asymptotic measure of extremal dependence which relies on Extreme Value Theory.

Extreme Value Theory: branch of statistics which studies the properties of extreme values and enable to fit them with theoretical distributions.
Strengths: robustness (large values + small samples); extrapolation (infer behaviour of tails from few rare extremes); inference and uncertainty (intrinsic with MLE and pdf); non-stationary fit (evolution of extremes with Climate Change).



Challenges of Extremes = Rarity + Magnitude:

- small sample + large values = large uncertainties! Need robust and resistant statistical approaches + inference.
- Scores exhibit statistics unstable behaviour, oversensitivity to bias, non-informative asymptotic limits.



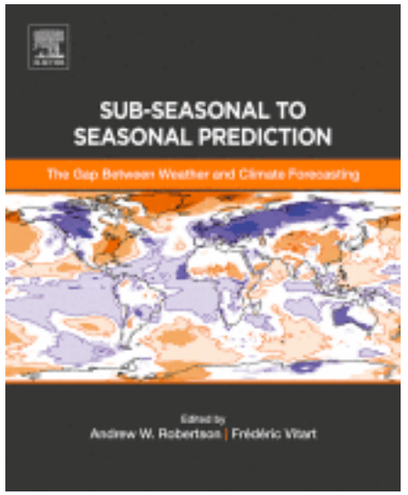
Coles (2001) "An introduction of Statistical Modelling of Extreme Values", Springer, 208 pp

Forecast verification activities

Caio Coelho

Overview of S2S verification methods and practices

Chapter 16: “Forecast verification for S2S time scales” by Caio A. S. Coelho, Barbara Brown, Laurie Wilson, Marion Mittermaier, Barbara Casati, in “Sub-seasonal to seasonal prediction: the gap between weather and Climate, (2019)” AW. Robertson and F.Vitart ed., Elsevir

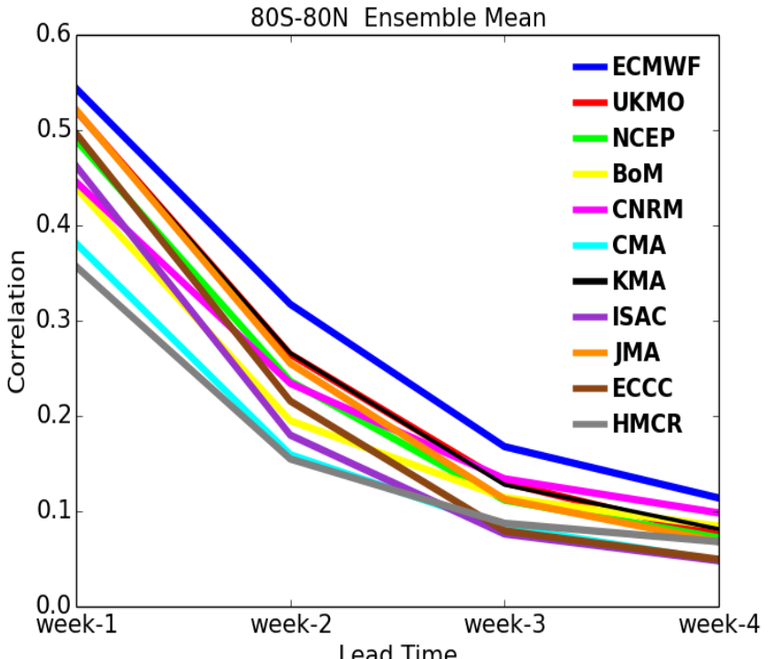


Proposed a verification framework for South American sub-seasonal precipitation predictions

Coelho, C. A. S.; M. A. F Firpo. F. M. de Andrade (2018) A verification framework for South American sub-seasonal precipitation predictions. Met.Zet. 27: 503-520.

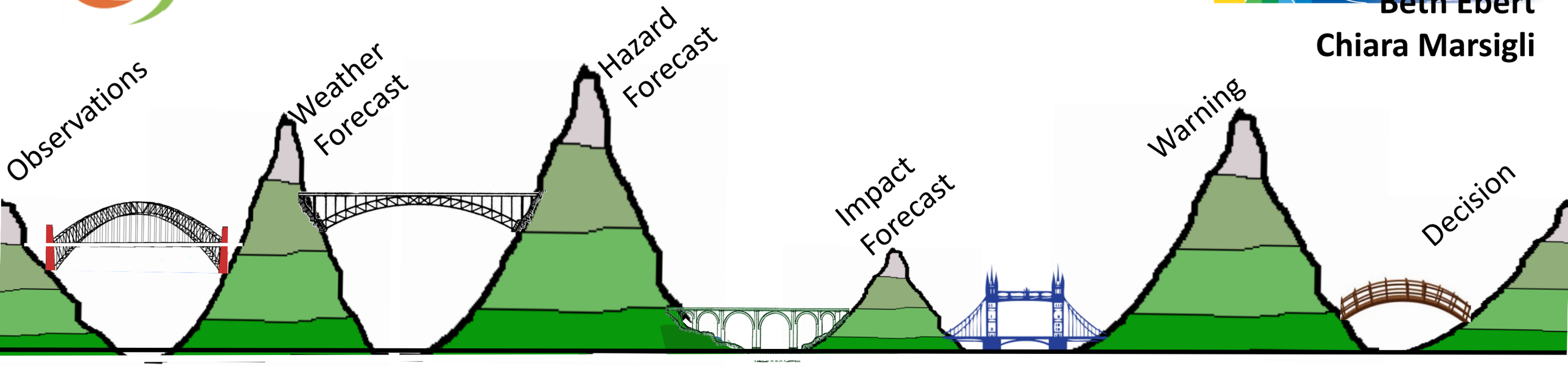
Performed global precipitation hindcast quality assessment of all S2S project models

de Andrade, F.M., Coelho, C.A.S. & Cavalcanti, I.F.A. (2019) Global precipitation hindcast quality assessment of the Subseasonal to Seasonal (S2S) prediction project models. Climate Dynamics 52, 5451–5475.



HIWeather and the User Value Chain

Beth Ebert
Chiara Marsigli



Sensor Technology | Weather Modelling | Ocean, Land Surface & Pollution Modelling | Social, Economic & Health Modelling | Communication Science: Messages & Media | Behavioural Psychology

Verification of HIW:

- Increased use of spatial verification (e.g. MODE)
- More frequent use of probabilistic verif. approaches
- Greater use of 3rd party observations, crowdsourcing:

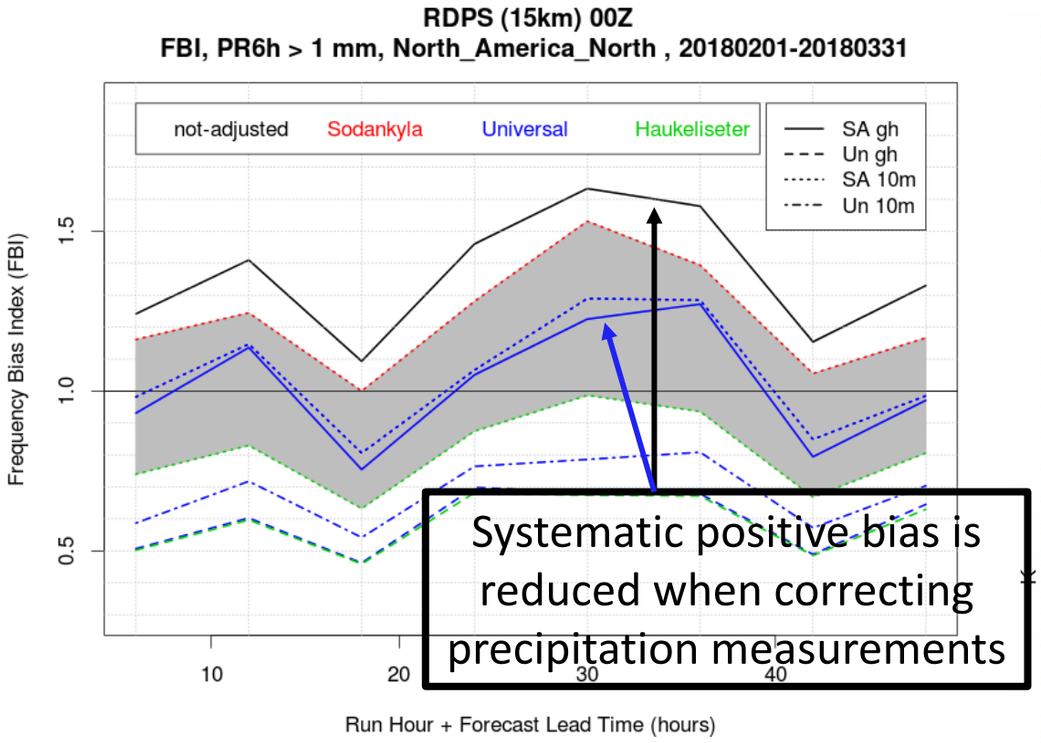
2nd International Verification Challenge

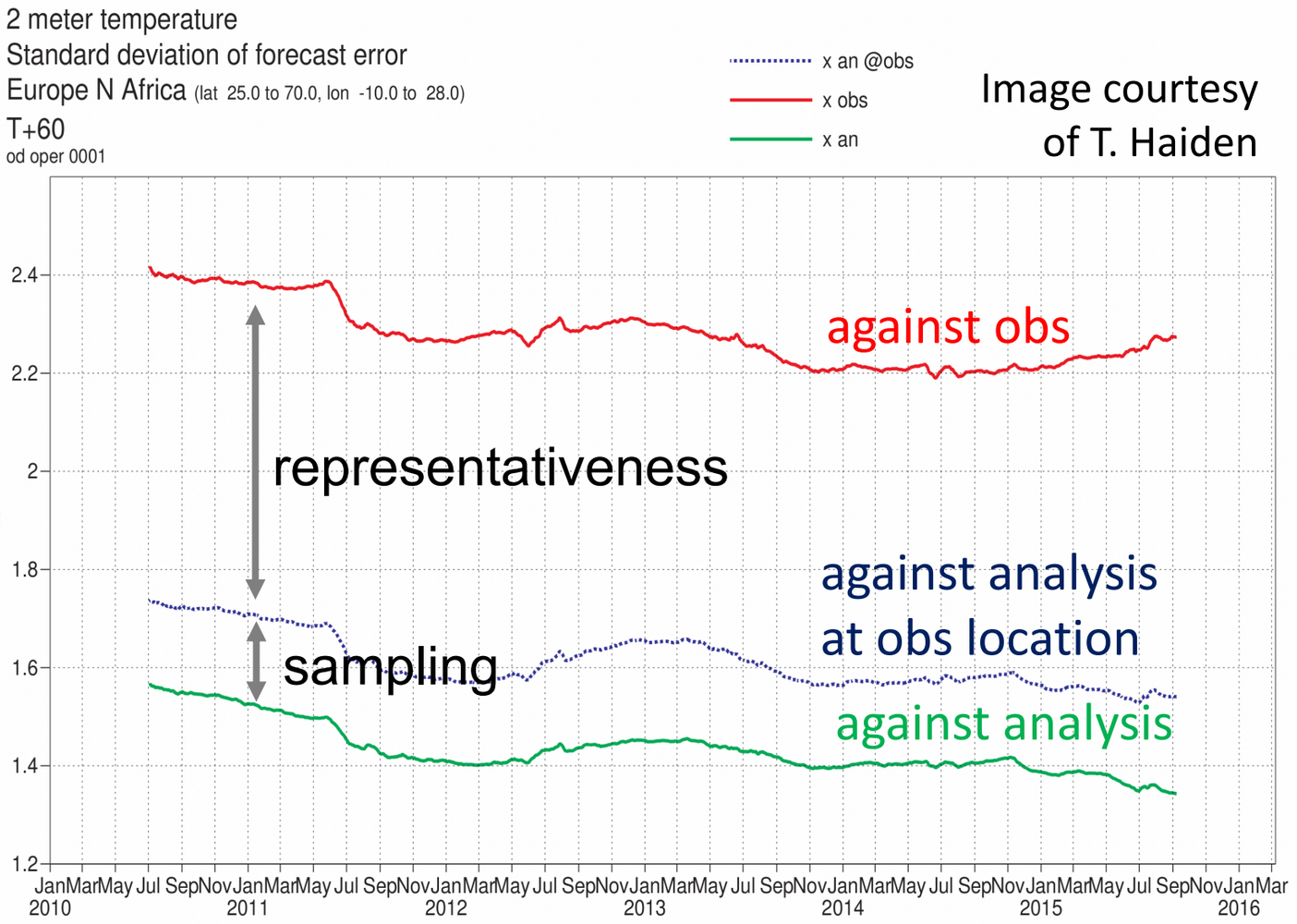
Challenges and new directions:

- The elephants in the verification room: observation uncertainty and representativeness => exploit DA techniques
- New observations : spatial, 3d satellites, vertical profiles, mobile networks …
- physical processes and interactions: conditional, multi-variate statistics;
- error tracking (linkages, teleconnections) and causality.
- Methods tailored to different research and user communities: sea-ice, ocean, hydrology, urban …

- Artificial Intelligence and Machine Learning: error detection, post-processing (provided the sample size is large enough to avoid overfitting)

Representativeness and observation uncertainty … can dominate the forecast error …



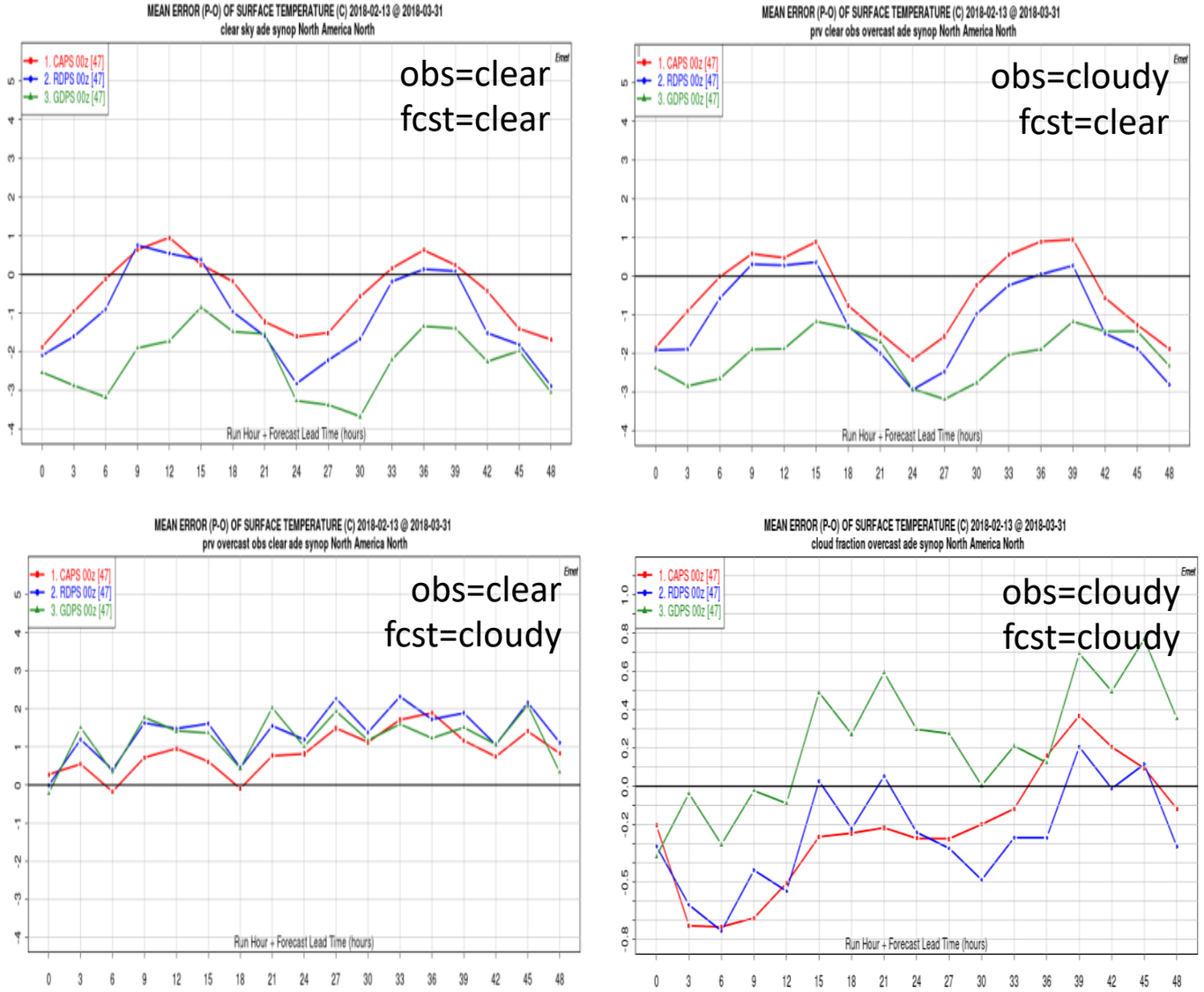


Obs error: Bowler (2008), Candille and Talagrand (2008), Ferro (2017).

Process Diagnostics

- Aim: provide feedback on the physical nature of the forecast error.
- Haiden et al (2019) outline some strategies for process diagnostics.
- **Processes: interactions between physical variables => conditional and multivariate**

2m Temperature bias
conditioned on cloud cover

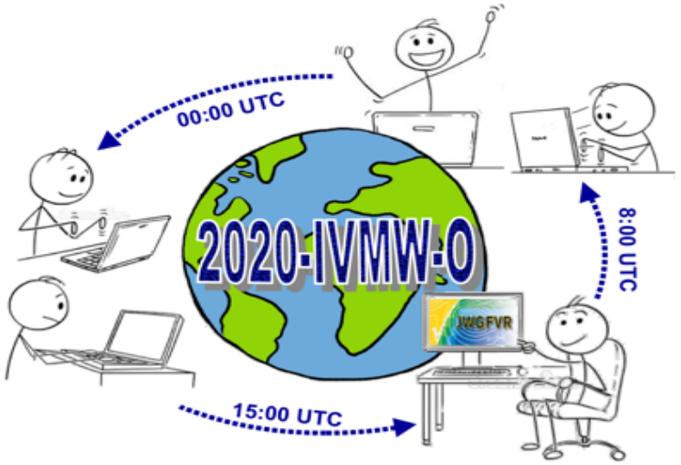


Opportunities:

Testbed datasets (e.g. YOPPsiteMIP): paired obs-forecast multiple variables.

ESM increasing complexity: **CMIP, WGNE, GEWEX intercomparison projects**

This workshop



Online around-the-clock format: a paradigm change!
More than 300 registered participants, very high quality submissions (we had to increase the sessions)!!

Highly international: winning institutions are
Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC),
Servicio Meteorológico Nacional **(**Argentina), Nigerian Met Agency,
further than CPTEC / INPE (Brazil), NCRMWF (India),
NOAA, ECCC, BoM, UK MetOffice, ECMWF.

Servicio Meteorológico Nacional

cmcc Centro Euro-Mediterraneo sui Cambiamenti Climatici

NiMet Nigerian Meteorological Agency

Outreach several new groups / scientific communities (s2s and climate, processes, polar and sea-ice, ocean ...): two-way exchange between verification experts in multiple disciplines.

Your contribution is shaping the future: Thank you!