Forecast Verification in the YOPP Framework

YOPP summit, WMO, Geneva, 13 July 2015

Talk outline:

1. Key factors for a succesful verification strategy
   - identified users
   - key variables and polar processes
   - spatial verification approaches
2. Verification strategy recommendations
   - model diagnostics (feedback to developers)
   - summary performance measures (monitor and compare)
   - user-oriented verif (transport sector)
3. Observation challenges

Barbara Casati

Environnement Canada
Environment Canada

Pertti Nurmi

Finnish Meteorological Institute

Members of the WWRP Joint Working Group on Forecast Verification Research (JWGFVR)

JWGFVR

Foreword: the role of the JWGFVR

Quote: WMO congress, cg17, 2015

Resolution 4.3(5)/2 on YOPP

Decides that WMO should support a period of intensive observing, numerical modelling simulations, **verification**, user-engagement, and education activities through the Year of Polar Prediction (YOPP), planned from mid-2017 to mid-2019, and a subsequent research consolidation phase in order to enable a significant improvement in environmental prediction capabilities for Polar Regions and beyond.

WG on Forecast Verification Research

4.3(5).26 Congress recognized the relevant activities of the Joint Working Group on Forecast Verification Research and encouraged WCRP, WWRP and GAW to further explore synergies for a seamless and consistent approach to forecast verification across scales and disciplines. Such an approach to forecast verification will address the needs of the CAS and WCRP projects and assist in the development of new verification methods required for operational services.

⇨ **Report with recommendations for YOPP verification activities**
(in preparation, to be included in next YOPP implementation plan)

1. Key factors to consider for a successfull verification strategy:

1. who are the verification end-users (e.g. modellers or end-users, such as navigation companies);
2. what are the verification purposes (e.g. diagnostics or administrative);
3. what are the questions to be addressed (e.g. model predictability limit) and/or the attributes of interest (e.g. timing for onset and clearance of fog);

4. the type of forecast verified (e.g. continuous, categorical or probabilistic);
5. which are the statistical characteristics of the variables to be verified (e.g. smooth upper-air variables, such as GZ and temperatures, or spatially episodic / discontinuous variables, such as precipitation or sea-ice).
6. which are the available observations (e.g. in-situ station obs or satellite-based spatial obs)

First three bullets, 1-3: properly formulate the verification question
Last three bullets, 4-6: technical aspects of the verification strategy

There is no single verification technique that can address all these!
Each verification strategy ought to be tailored to the user needs and their verification purposes and questions, and to the forecasts and variables verified, and the corresponding available observations.

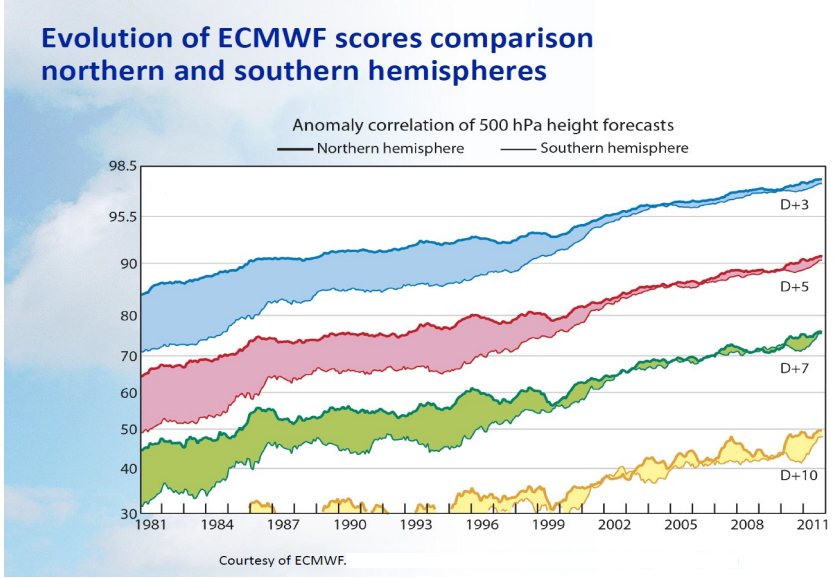
For recommendations on the YOPP verification strategy we have identified three classes of users / verification purposes:

1. Diagnostics for model developers: identify sources of error; compare different model schemes.

2. Summary verification scores for administrative / generic purposes (compare and monitor progresses).

3. Meaningful verification measures for selected end-users:

- focus on the **transport sector** (aviation, marine and land transport).
- address the verification of environmental variables such as **sea-ice**, snow, permafrost, ceiling and icing, fog and visibility.



Key variables and polar processes

Basic (surface and upper-air) atmospheric variables: temperature, dew-point temperature, precipitation, cloud cover, relative humidity, wind speed and direction, geopotential height, mean sea level pressure.

Environment surface variables: Sea-ice. Snow at the surface (snow cover, snow thickness). Permafrost (soil temperature).

Modeling challenging processes / variables:

- coupling atmosphere - surface - ocean -cryosphere.
- surface-atmosphere exchanges (turbolence, energy and momentum fluxes, radiation budget).
- stable boundary layer representation (temperature verical profile).
- effects of steep orography.
- clouds, mixed phase clouds.

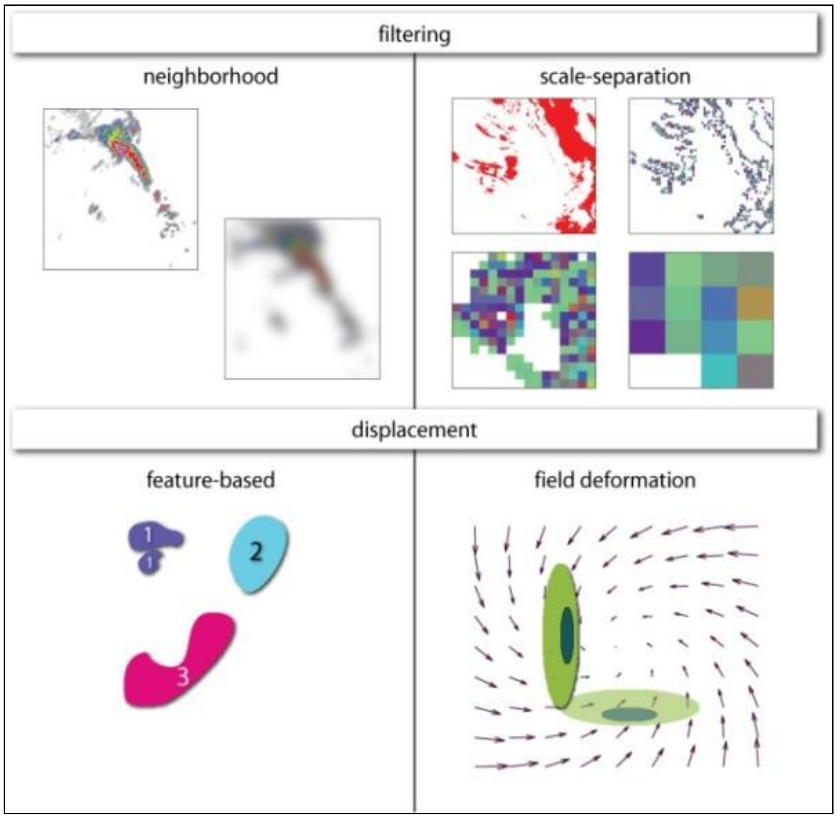
High Impact Weather: polar lows, low-level jets, topographically influenced flows such as katabatic winds and hydraulic shocks, extreme thermal contrasts, blizzards, freezing rain, fog → Collaboration and interplay with verification activities of the WWRP High Impact Weather (HIW) project.

User-relevant variables: visibility, ceiling and icing (for aviation); sea-ice, fog and visibility (for navigation); ground conditions (e.g. snow, permafrost) for land transport.

Spatial verification approaches

- account for **coherent spatial structure** and the presence of **features**
- provide information on **error in physical terms (meaningful verification)**
- assess **location and timing errors** (separate from **intensity error**)
- account for **small time-space uncertainties** (avoid **double-penalty** issue)

Neighborhood: relax requirement of exact space-time matching



Scale-separation: analyse scale-dependency of forecast error

Feature-based: evaluate attributes of isolated features

Field-deformation: use a vector and scalar field to morph forecast into obs

From Gilleland et al 2010

MesoVICT: inter-comparison of spatial verification methods
http://www.ral.ucar.edu/projects/icp/

2. YOPP Verification Strategy Reccomendations

2.1 Model Diagnostics

- Process-based diagnostic verification to provide feedback to modelers.
- Multi-variate obs (embrace all physical variables/aspects of the process) at high resolution and frequency (super-sites).

  ➔ *Model diagnostics are very specific to the physical process analyzed and ought to be set-up with modellers*

- **Simple yet informative statistics, informative graphical display**
  e.g. Hovmoller detect flow-dependent error propagation
- **Spatial verification approaches**
  e.g. Jung and Leutbecher (2008)

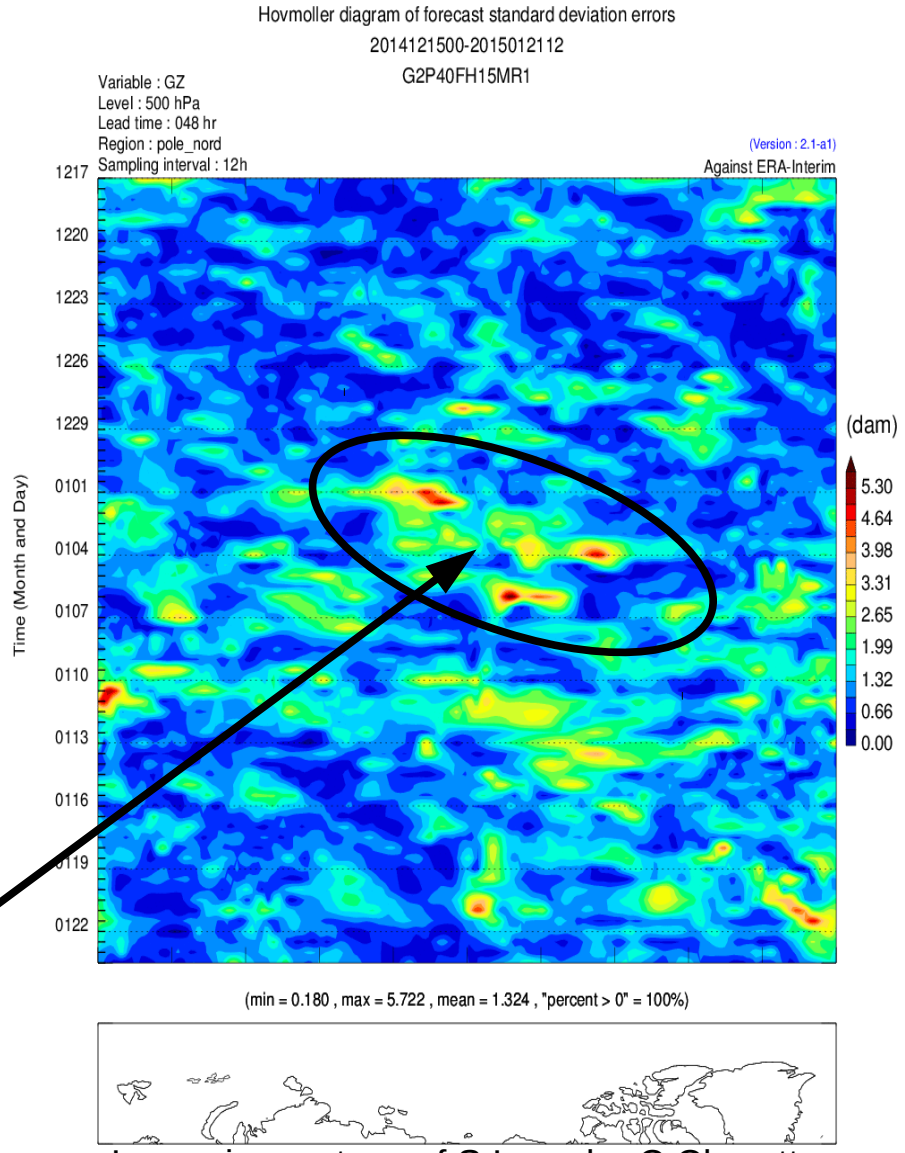
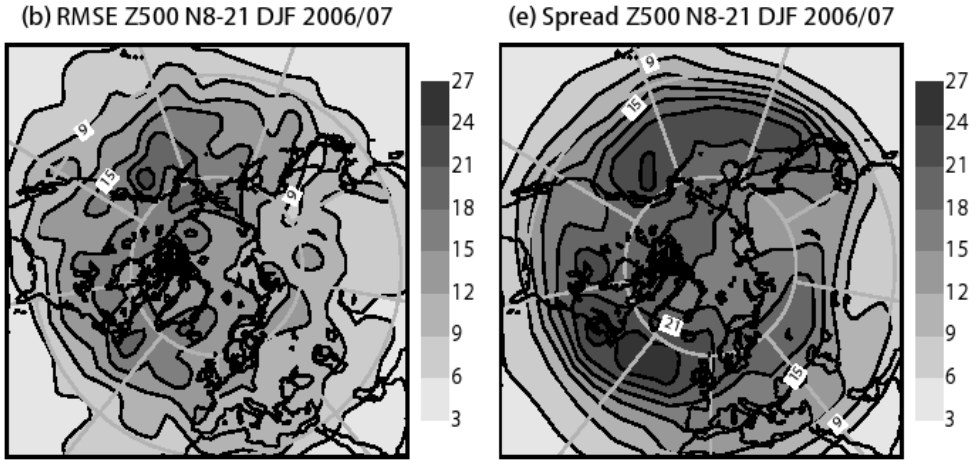


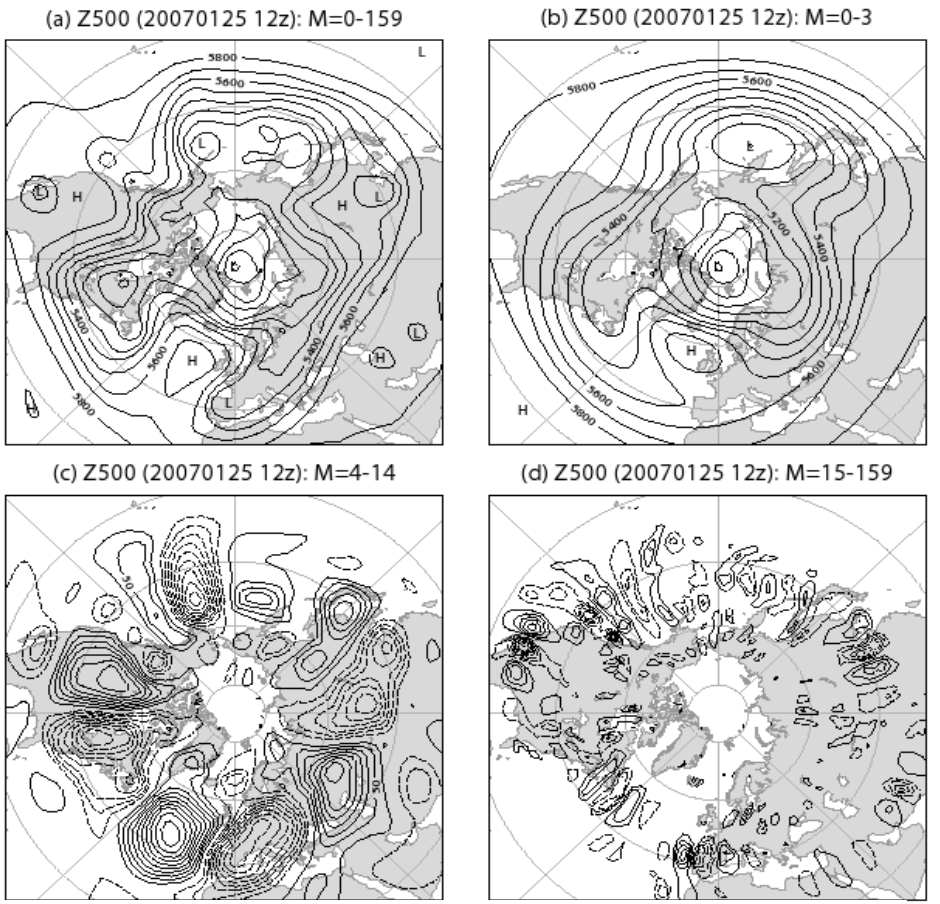
Image is courtesy of S.Laroche C.Charette

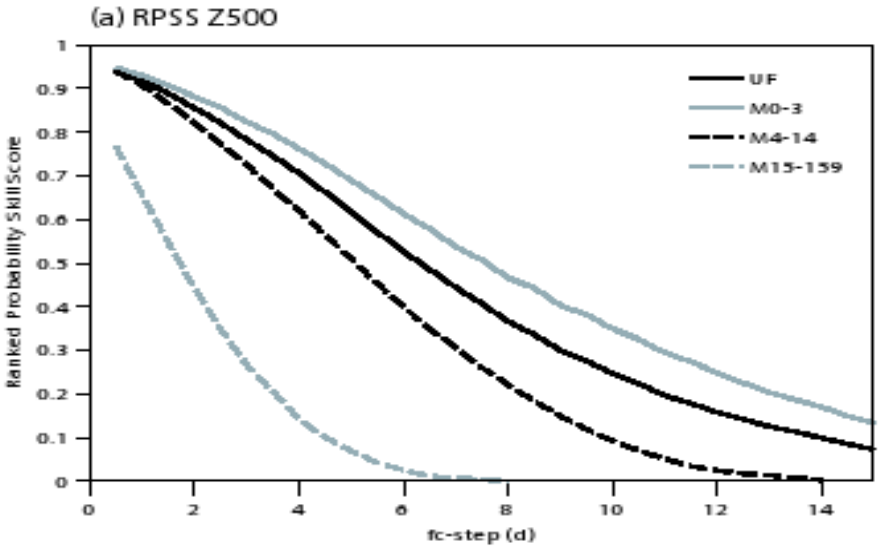
Jung and Leutbecher (2008): Scale-dependent verification of ensemble forecasts. QJRMS 134

1. Spherical harmonics: separate planetary, synoptic and sub-synoptic scales.
2. Evaluate skill on different scales (BSS, RPSS).
3. Analyse the scale dependency of the spread-skill relationship.



Ensemble is over-dispersive at the synoptic scales. Max spread and error correspond the North Atlantic and North Pacific storm track regions



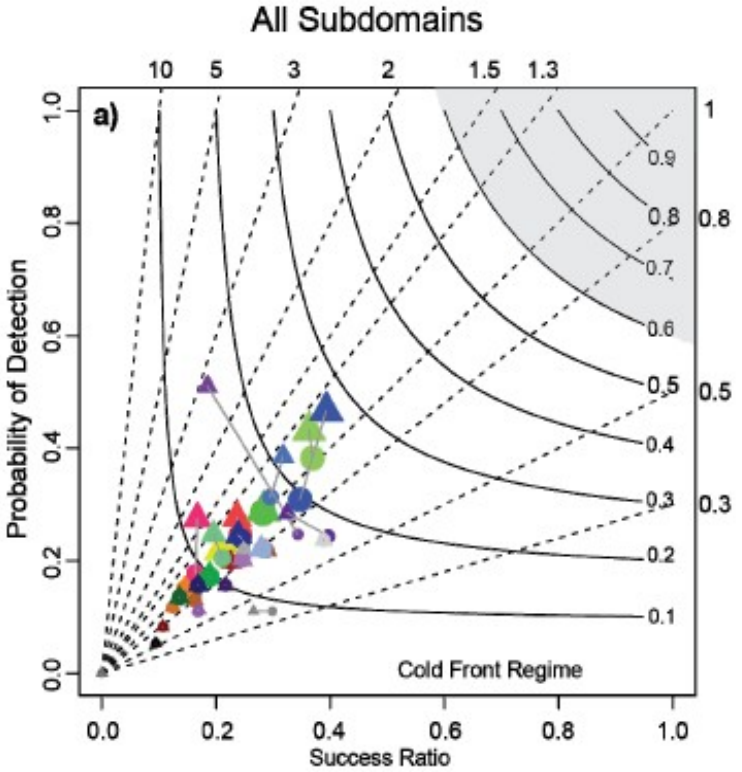


2. YOPP Verification Strategy Reccomendations

2.2 Summary verification scores

Verif of **basic surface and upper-air** atmospheric variables should meet **CBS standards**:

- **Continuous scores** for Gaussian and continuous variables (GZ, temperature): bias, MSE, MAE, MSE SS, correlation, S1, …
- **Categorical scores** for right-skewed episodic / discontinuous variables (precip, clouds): FBI, TS and ETS, PC, HSS, OR, YQ, …
- Compare multiple scores on **summary performance diagrams: Taylor (2001), Roebber (2009).**
- Traditional scores degenerate to un-informative trivial values as the events become rarer: **Extreme Dependence Indices (Ferro and Stephenson 2011).**



Example from Roberts et al. 2011, after Roebber 2009 and C. Wilson 2008
Dotted: BIAS isolines
Continuous: TS isolines

- **Ensemble and prob forecasts**: Brier score and SS, CRPS (res+rel+unc), ROC and reliability diagrams, Talagrand histogram, error-spread relationship. See TIGGE museum (Dr. M. Matsueda) **http://gpvjma.ccs.hpcc.jp/TIGGE**

Statistical significance

Verification practices need to handle the delicate balance between:

- Multiple cases **aggregation** → statistical significance
- **Stratification / conditional verification** → target specific process

Providing the statistical significance of the verification results is **fundamental**

Statistical significance can be performed either by **traditional parametric tests** or by using **resampling approaches, permutation tests and bootstrapping**

References: Wilks 2006, chapter 5; von Storch and Zwiers, 1999, chapter 6; Jolliffe 2007; Efron and Tibshirani, 1993.

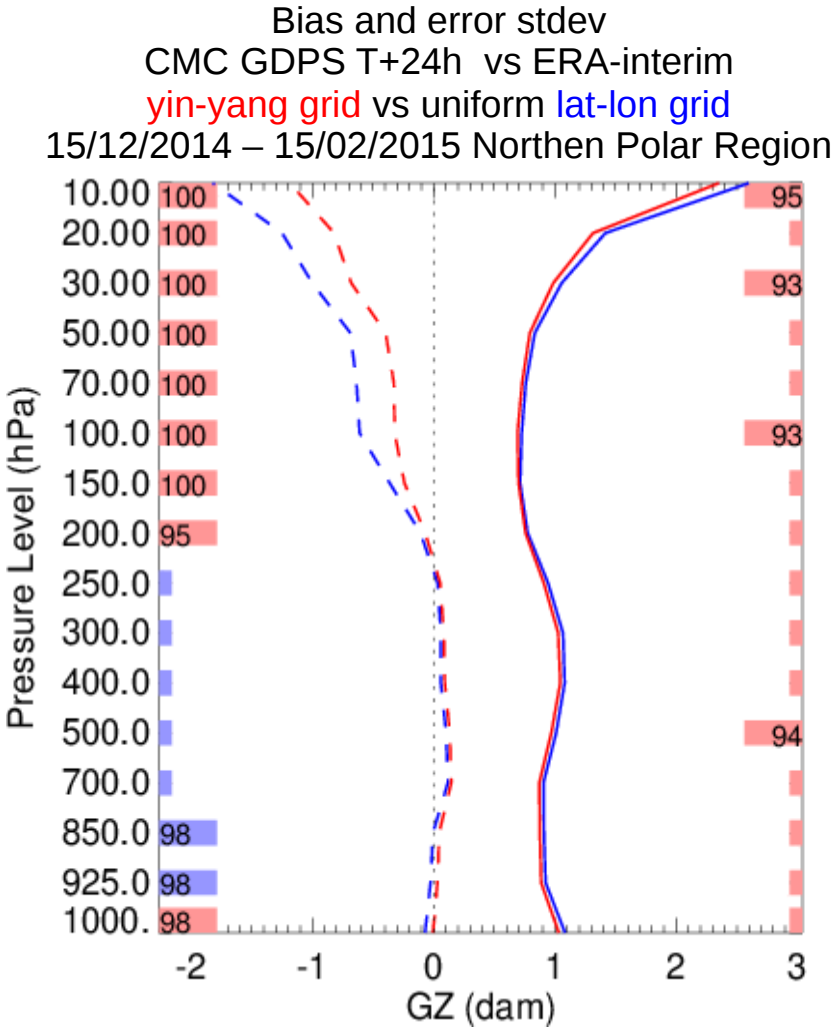


Image is courtesy of S.Laroche C.Charette

2. YOPP Verification Strategy Reccomendations

2.3 verification for selected end-users

Example: focus on sea-ice verification for navigation safety (Canadian Ice Service).

Several **attributes** to consider:

- Ice concentration
- Ice extent
- Ice edge
- Ice stage of development (age)
- Ice thickness
- Ice pressure
- Ice drift trajectories
- Iceberg tracking

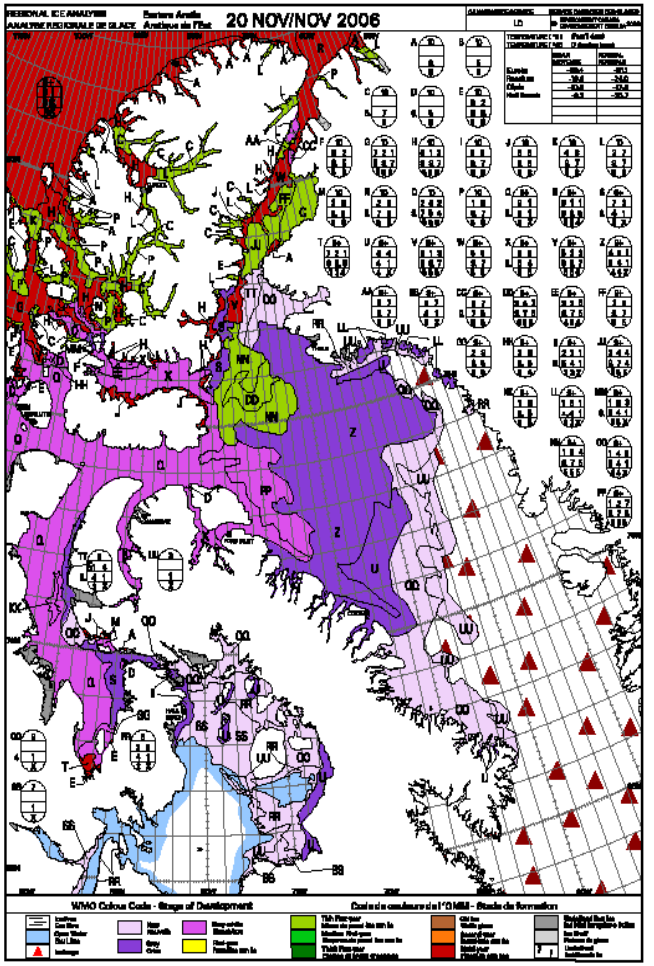
Several **issues** to overcome:

traditional scores limited in Marginal Ice Zone. threshold issues: multicategory scores?

Crucial for navigation!! Ship obs report solely with ice pressure: infer ice pressure from ice motion?

...

CIS ice chart



The more sophisticated the user, the more complex the problem: user-specific attributes / issues, need tailored verification approach!

Meaningful Verification

Distance to ice-edge: use a Partial Hausdorff Metric (the median distance). Intuitive verification statistics: provide a **distance in km!**

RIPS forecast, persistence and analysis vs IMS analysis for the whole 2011

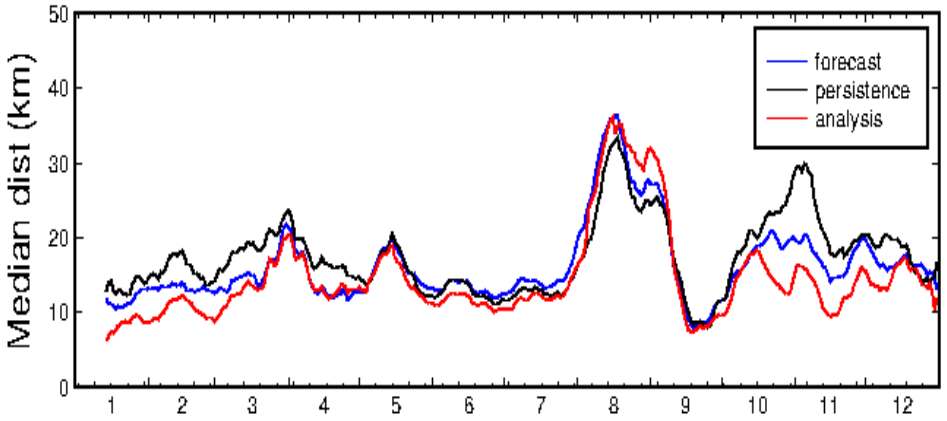


Image is courtesy of JF Lemieux (MRD/EC)

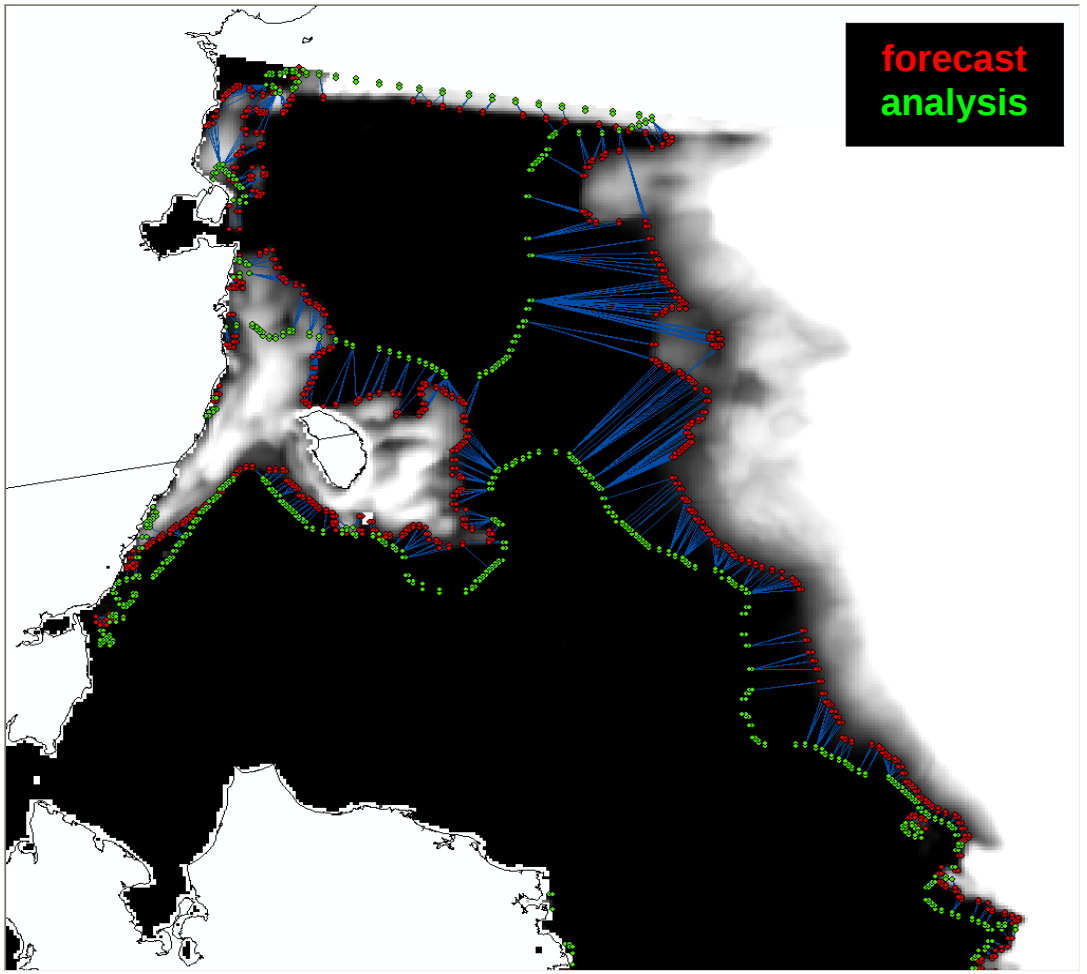


Image is courtesy of A. Cheng (CIS)

Sea-ice drift trajectories

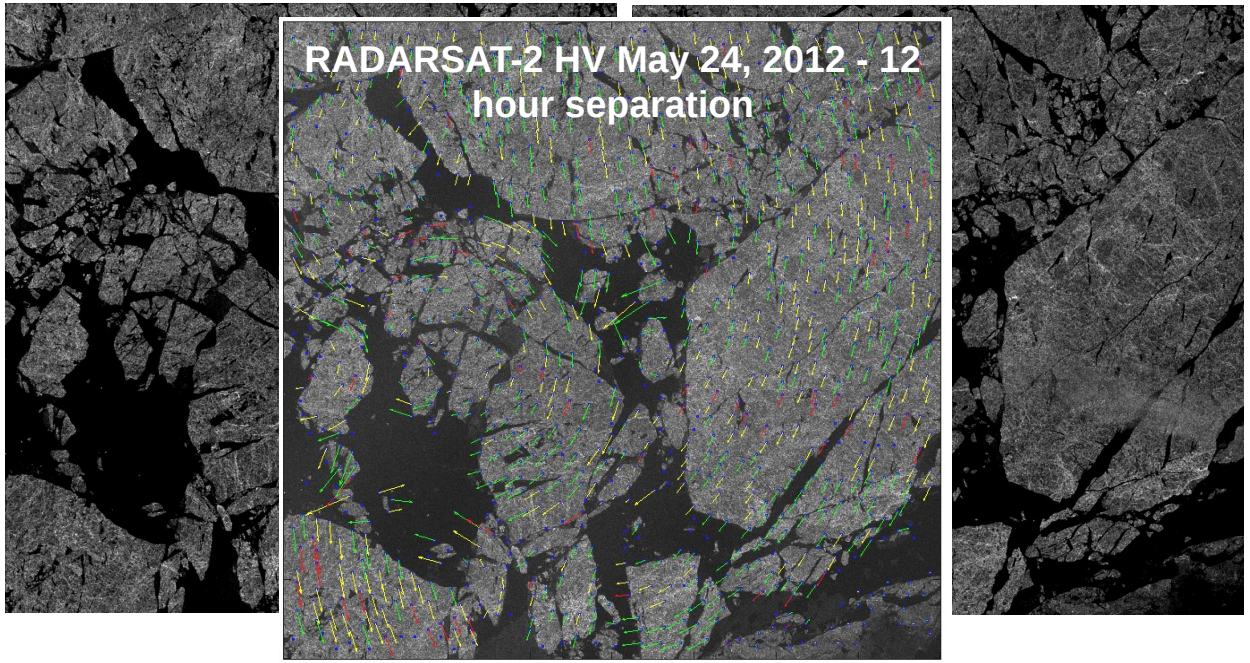
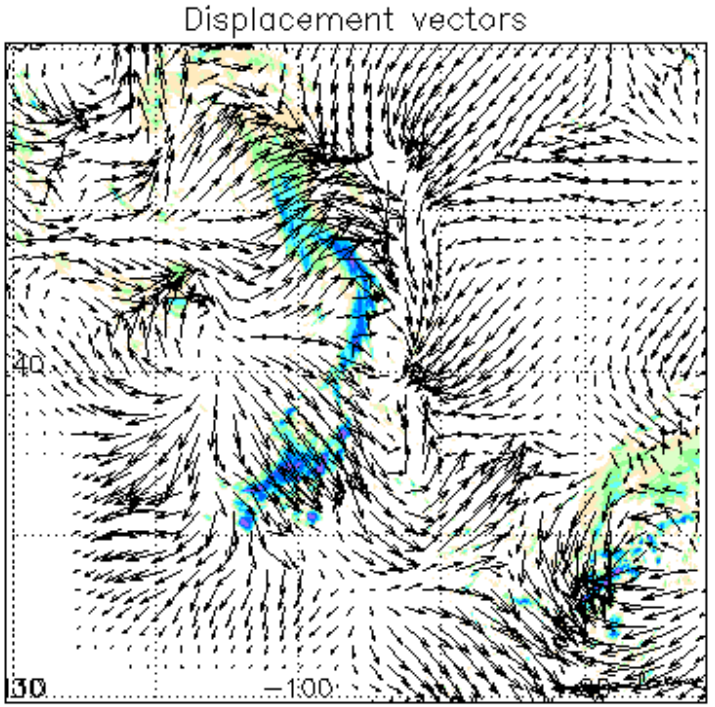


Image is courtesy of A. Cheng (CIS)



From Keil and Craig (2007, 2009)

The automated **sea-ice tracking** system (Komarov and Barber, 2014) produces a vector field similar to that produced by a **field-deformation verification approach:** verify sea-ice motion spatially.

Alternative: use the **S1 score** to verify modelled versus obs sea-ice motion field (the S1 score was historically designed to assess the accurancy of the forecast in reproducing the gradients of pressure or geopotential height, in consideration of the relationship of these gradients to the wind field).

Similarly: **feature-based** technique for **iceberg tracking**.

3. Observation Challenges

Obs at single sites:

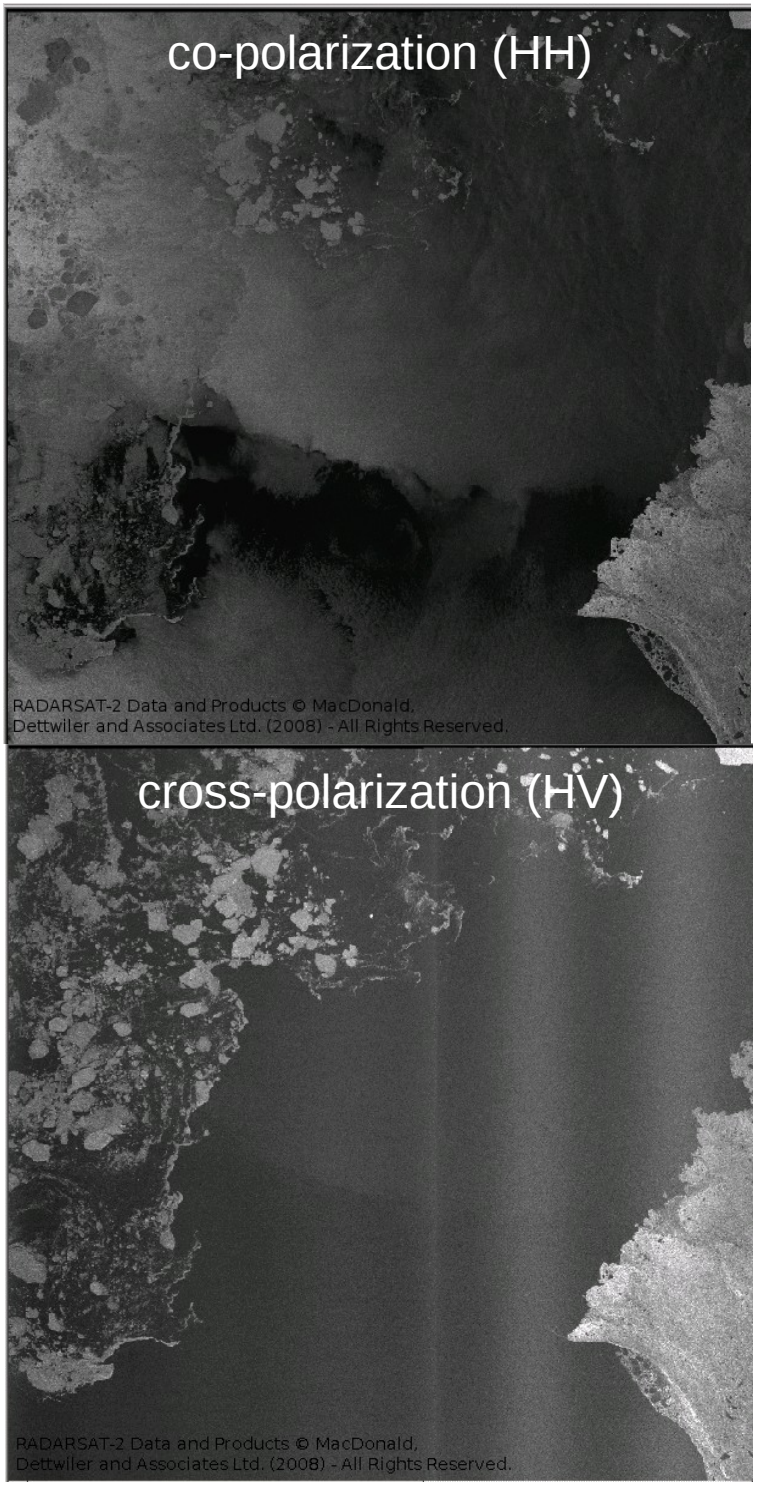
- Sparseness and inhomogeneity of surface observations.
- Station poor space-time representativeness (coastal, seasonality).
- Instrument failure.

Satellite-based gridded obs products:

Challenges: in polar regions the constant prevailing low stratus clouds hinders the use of satellite data. Verification activities will benefit from improvements of satellite data retrievals.

Advantages: enhance obs spatial cover; detect spatial patterns; spatial verification approaches; more informative graphical displays (e.g. Hovmoller diagrams, zonally/meridionally avg versus lead time or vertical profile).

Disadvantages: assumptions and uncertainties associated to remote sensing and to the gridding process.



Images are courtesy of A. Cheng (CIS)

Should we perform

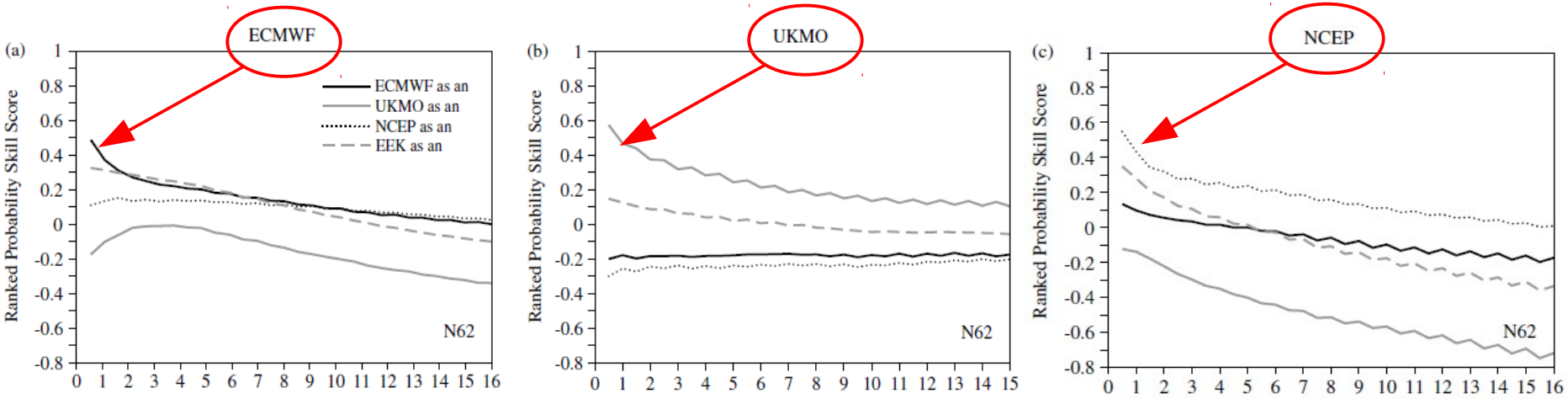
verification against analyses ?

Verification against model-based analysis **can be informative**, but must be performed with **awareness** and acknowledging its **drawbacks**.

Advantages: 1. obs quality control, ingestion of the obs uncertainty, and represenattiveness issue are dealt within the gridding process; 2. obs is spatially defined (covers the whole space-time domain) → spatial verif approaches, informative graphical display.
Disadvantages: uncertainties / assumptions for quality control and gridding.

Model-based analysis might reject good observations (polar regions are more affected than mid-latitudes, due to lack of buddy-check obs and large model biases). Good practice is to perform verification against analysis solely in concurrence of recently assimilated observations (Lemieux et al, 2015).

Verification against own analysis leads to best score (Park et al, 2008)



RPSS for 850 hPa temperature in the tropics, TIGGE data (from Park et al, 2008, figure 14)

Handeling observation uncertainties in verification practices

Solely few studies **account for obs uncertainties / sparseness in the verification approach / scoring algorithm.** Ciach and Krakewski (1999), Bowler (2008), Santos and Ghelli (2011), Mittermaier (2014), Casati et al (2014): **future verification research focus!**

Awareness of assumptions and weaknesses of (gridded) obs dataset in the interpretation of verification results. Could use a model-to-observation approach: e.g. use model-simulated radiances for a comparison against satellite observed radiances, directly.

Verification against multiple gridded datasets (and/or analyses) is reccomended: the uncertainty / spread between analyses / gridded observation datasets should be (an order of magnitude) smaller than the forecast error. Some verification statistics (e.g. Brier Score, CRPS, KS distance), can directly compare the distributions of and ensemble of models versus an ensemble of analyses.

Conclusions

- **Verification strategies ought to be tailored** to the user needs and verification purposes, and to the forecasts, variables verified, and the corresponding available observations.

    1. Diagnostics for model developers (error sources)
    2. Summary performance measures (monitor, compare)
    3. Meaningful verification for selected end-user

- YOPP: opportunity to exploit some of the new **spatial verification** approaches (alongside with traditional verification).
- Handle **obs sparseness / uncertainties**: verification research focus! Verification against multiple (gridded) dataset is encouraged.
- Well structured common **data archive** is crucial for YOPP verification activities.
- Who are the verifiers?

THANK YOU!

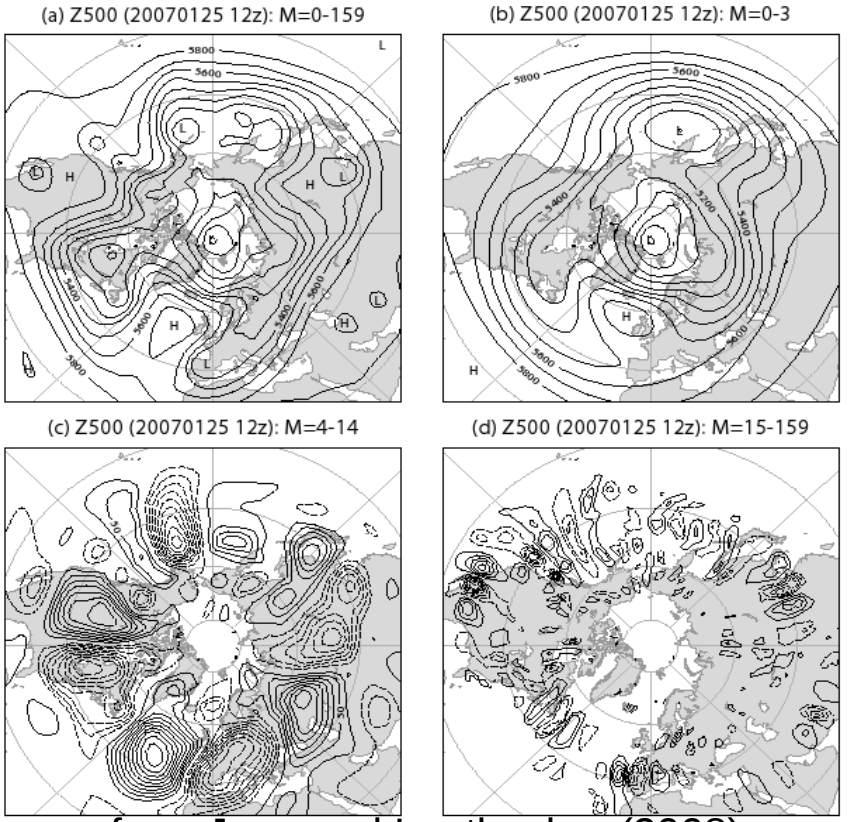
barbara.casati@ec.gc.ca

Extra Material

1. Scale-separation approaches

Briggs and Levine (1997), wavelet cont (MSE, corr);
Casati et al. (2004), Casati (2010), wavelet cat (HSS, FBI, scale structure)
Zepeda-Arce et al. (2000), Harris et al. (2001), Tustison et al. (2003), scale invariants parameters;
Casati and Wilson (2007), wavelet prob (BSS=BSSres-BSSrel, En2 bias, scale structure);
Jung and Leutbecher (2008), spherical harmonics, prob (EPS spread-error, BSS, RPSS);
Denis et al. (2002,2003), De Elia et al. (2002), discrete cosine transform, taylor diag;
Livina et al (2008), wavelet coefficient score. **De Sales and Xue (2010)**

1. Decompose forecast and observation fields into the sum of spatial components on different scales (wavelets, Fourier, DCT)
2. Perform verification on different scale components, separately (cont. scores; categ. approaches; probability verif. scores)

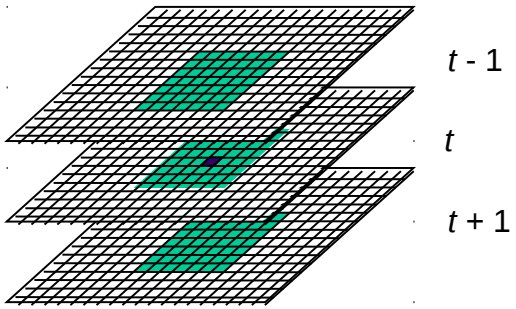


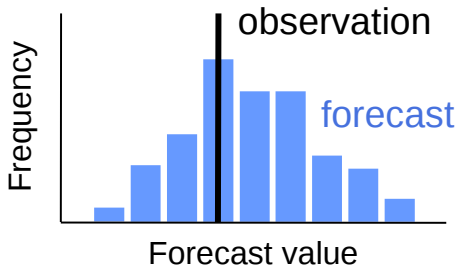
from Jung and Leutbecher (2008)

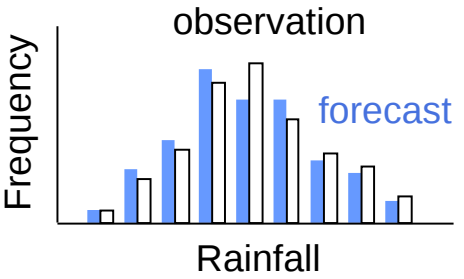
➔ Assess scale structure
➔ Bias, error and skill on different scales
➔ Scale dependency of forecast predictability (no-skill to skill transition scale)

2. Neighbourhood verification

1. Define neighbourhood of grid-points: relax requirements for exact positioning (mitigate double penalty: suitable for high resolution models); account for forecast and obs time-space uncertainty.

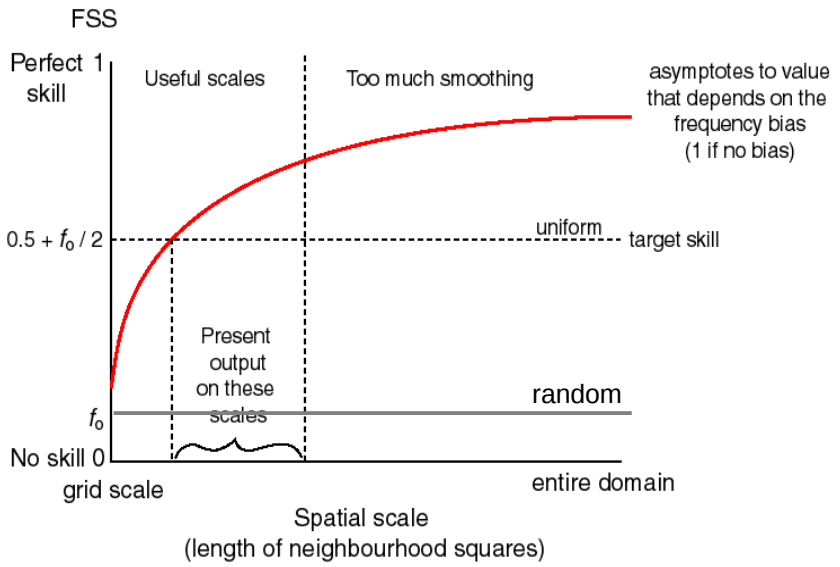






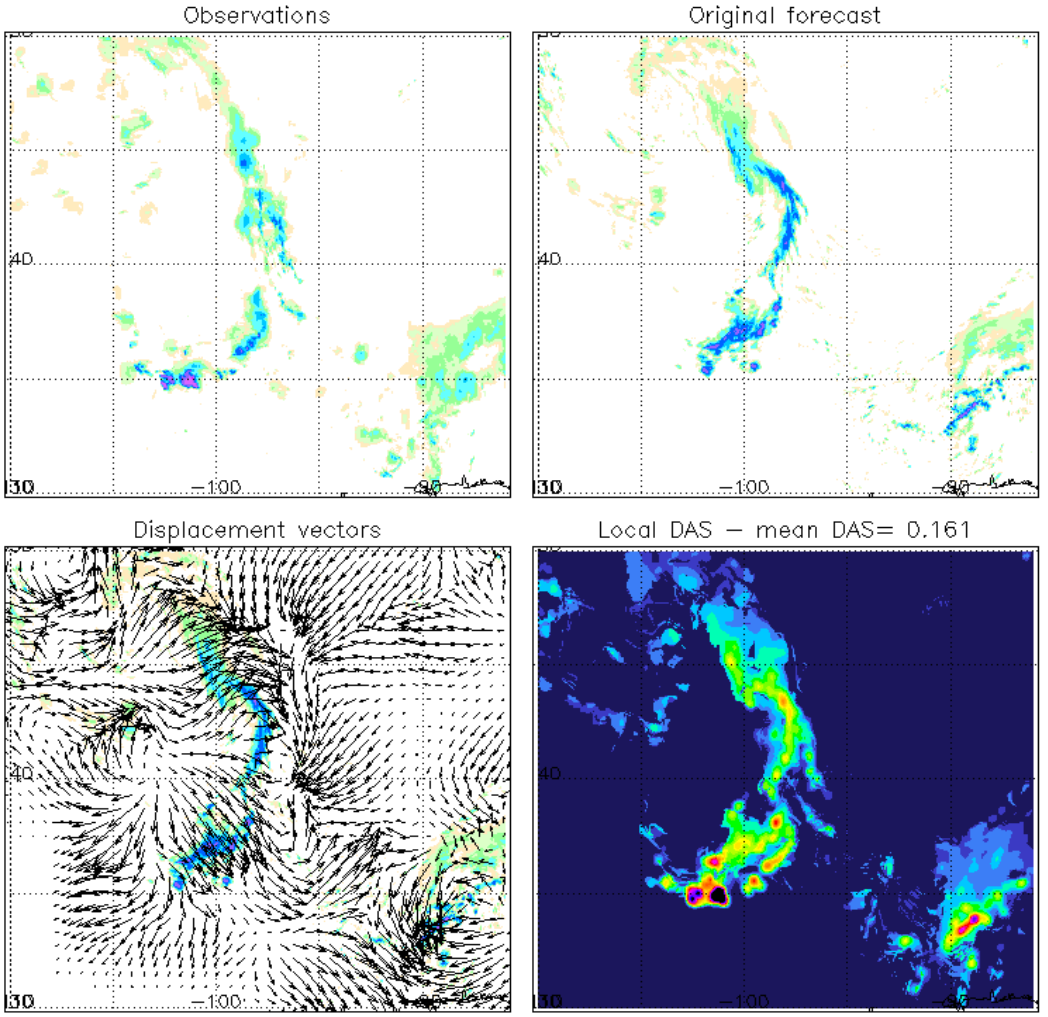
2. Perform verification over neighbourhoods of different sizes: verify deterministic forecast with probabilistuic approach

Yates (2006), upscaling, cont&cat scores; **Tremblay et al. (1996)**, distance-dependent POD, POFD; **Rezacova and Sokol (2005)**, rank RMSE; **Roberts and Lean (2008)** Fraction Skill Score; **Theis et al (2005)**; pragmatical approach; **Atger (2001),** spatial multi-event ROC curve; **Marsigli et al (2005, 2006)** probabilistic approach.



3. Field-deformation approaches

Hoffmann et al (1995); Hoffman and Grassotti (1996), Nehrkorn et al. (2003); **Brill (2002); Germann and Zawadzki (2002, 2004); Keil and Craig (2007, 2009)** DAS; **Marzbar and Sandgathe (2010)** optical flow; **Alexander et al (1999), Gilleland et al (2010)** image warping

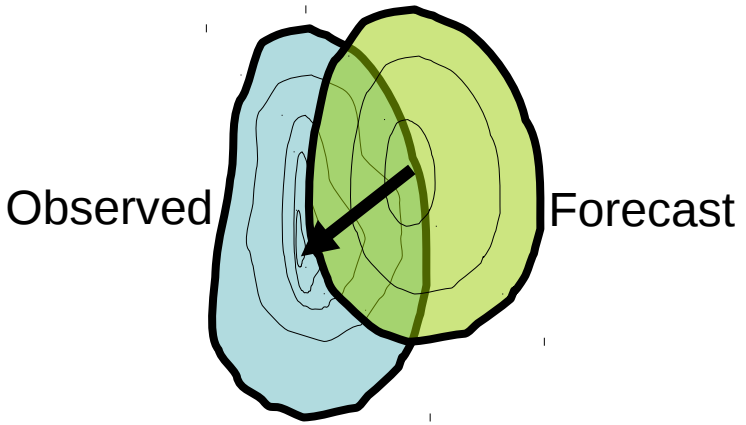


1. Use a vector (wind) field to deform the forecast field towards the obs field
2. Use an amplitude field to correct intensities of (deformed) forecast field to those of the obs field

- Vector and amplitude fields provide physically meaningful diagnostic information: feedback for data assimilation and now-casting.
- Error decomposition is performed on different spectral components: directly inform about small scales uncertainty versus large scale errors.

4. Feature-based techniques

- **Ebert and McBride (2000), Grams et al (2006), Ebert and Gallus (2009)**: CRA
- **Davis, Brown, Bullok (2006) I and II, Davis et al (2009):** MODE
- **Wernli, Paulat, Frei (2008):** SAL score
- **Nachamkin (2004, 2005):** composites
- **Marzban and Sandgathe (2006):** cluster
- **Lack et al (2010):** procrustes



1. Identify and isolate (precipitation) **features** in forecast and observation fields (thresholding, image processing, composites, cluster analysis)

2. assess displacement and amount (extent and intensity) error for each pairs of obs and forecast features; identify and verify attributes of object pairs (e.g. intensity, area, centroid location); evaluate distance-based contingency tables and categorical scores; perform verification as function of feature size (scale); add time dimension for the assessement of the timing error of precipitation systems.

New avenues for sea-ice verification

Variables:

- Ice concentration
- Ice extent
- Ice edge
- Ice stage of development (age)
- Ice thickness
- Ice pressure
- Ice drift trajectories
- Iceberg tracking

Canadian Ice Service:
Navigation (safety); Environment.

Ice concentration

Ice Concentration is measured in 10th: 0/10 = water; 10/10 = ice covered. It is a continuous variable, U-shaped distribution. It is characterized by a spatial coherent structure, spatial discontinuities, presence of features.

Marginal Ice Zone: transition between water and ice (where the action is)! For meaningful statistics, usually restrict the verification to the Marginal Ice Zone, e.g. evaluate verif statistics solely for grid-boxes where either obs/analysis and/or model have changed wrt previous day/week (e.g. van Woert et al 2004). Naturally compare to **persistence**.

Traditional verification approaches:
Continuous statistics (e.g. RMSE, Bias). Categorical scores (e.g. percent correct, misses, false alarms, frequency bias) from 2x2 contingency tables (e.g. Lemieux et al 2015). The latter implies thresholding.

Thresholding: continuous variable to categorical (water / ice).
Define **ice extent** and **ice edge**. We want to verify these spatially, with a physically meaningful approach.

Issues associated with the thresholding of sea-ice concentration

The natural threshold used to distinguish between ice / no-ice can be different in obs-based products versus model output.

Example (Smith et al 2015): IMS th = 0.4; GIOPS th = 0.2 (sea-ice concentration > 0.2 is associated with freezing SST, whereas sea-ice concentration < 0.2 is associated with no freezing SST).

Suggested (traditional) verification approach to (partially) address this issue: **multicategorical contingency table**.

Use different thresholds and define a scoring matrix which balances-out rewards and penalties, in order to accomodate the different-user perspectives.

Multi-category contingency table

Observed ice concentration

Predicted ice concentration

P(o1,f1)	*P(o2,f1)*	*P(o3,f1)*	*P(f1)*
P(o1,f2)	*P(o2,f2)*	*P(o3,f2)*	*P(f2)*
P(o1,f3)	*P(o2,f3)*	*P(o3,f3)*	*P(f3)*
P(o1)	*P(o2)*	*P(o3)*	1

$$FBIi = p(oi) / p(fi)$$

$$PODi = p(oi,fi) / p(oi)$$

$$PC = \Sigma\, p(oi,fi)$$

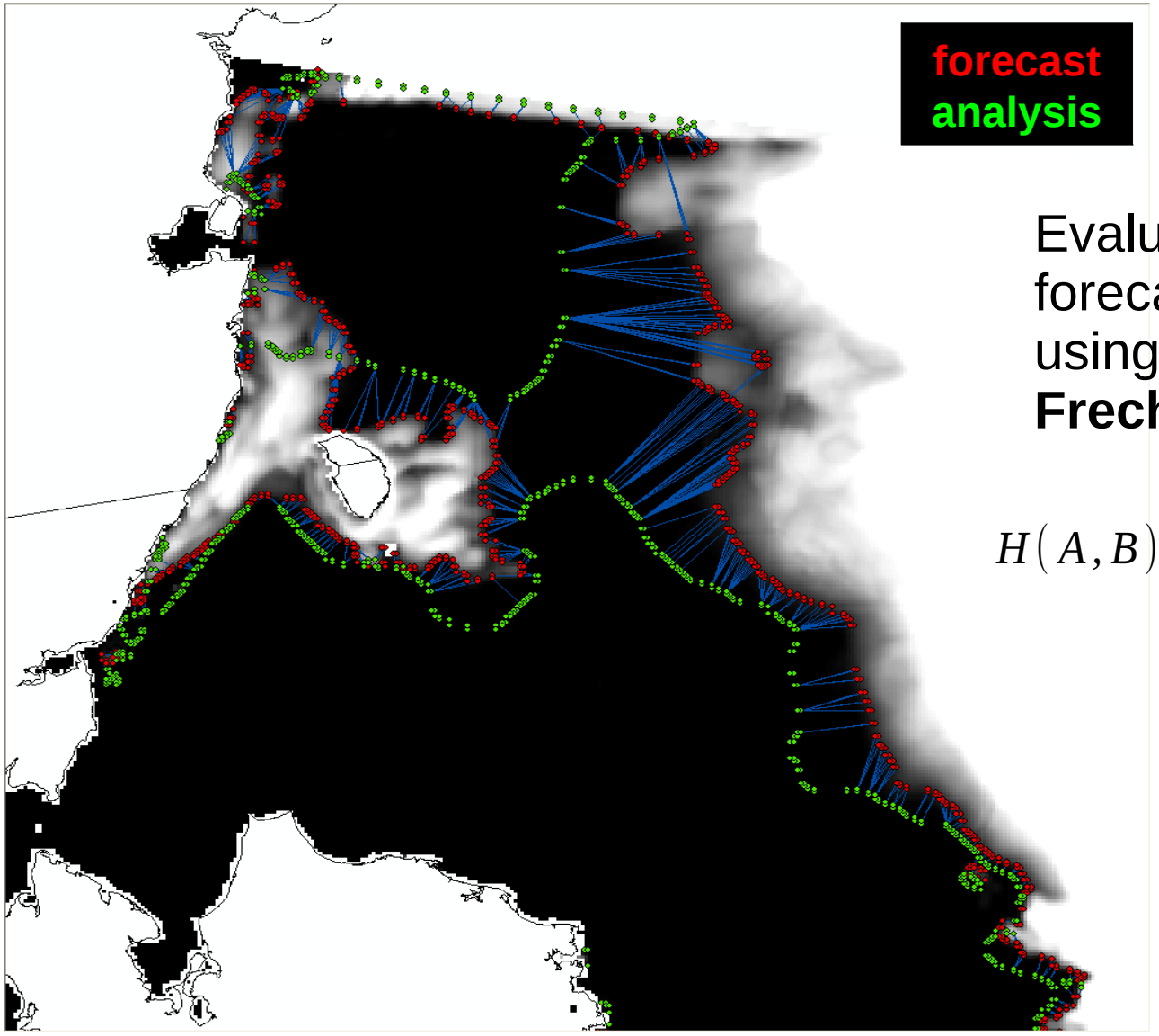
$$HSS = \frac{\Sigma\, p(oi,fi) - \Sigma\, p(oi)p(fi)}{1 - \Sigma\, p(oi)p(fi)}$$

Gerrity (1992) MWR 120, **Gandin and Murphy (1992)** MWR 120.

Develop a family of equitable scores. These are obtained by assigning to each joint probability p(oi,fi) a weight sij, obtained from meta-verification arguments. The **scoring matrix [sij]** is a tabulation of reward or penalty for every couple (oi,fj). The skill score is given by

$$SS = \Sigma\Sigma\, p(oi,fi)sij$$

Ice-edge verification



Courtesy of Angela Cheng (CIS)

Evaluate the distance between forecast and obs ice-edge by using the **Hausdorff** and **Frechet distances**

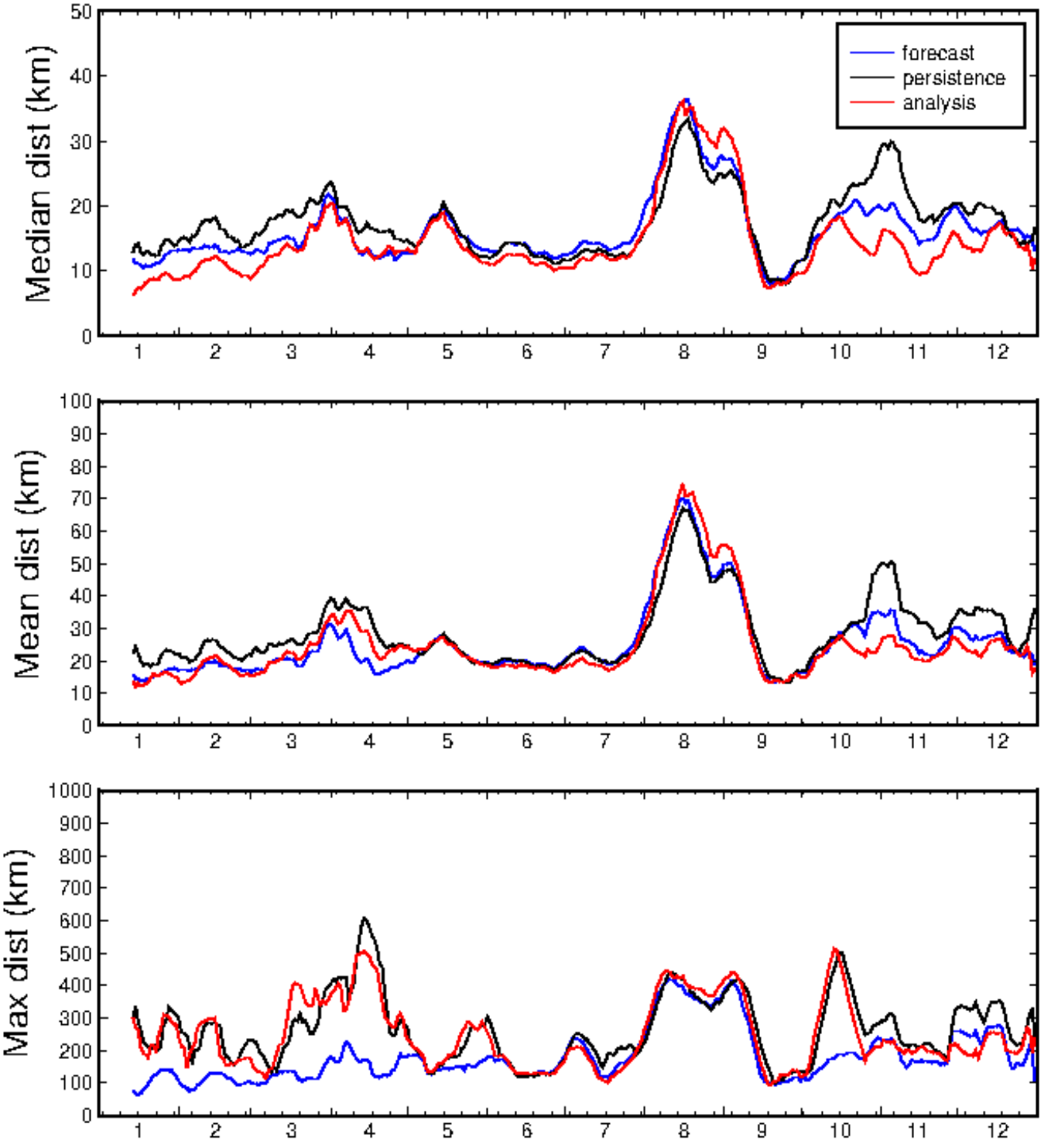
$$H(A,B)=\max\left\{\sup_{a\in A}\left(d(a,B)\right),\sup_{b\in B}\left(d(b,A)\right)\right\}$$

Meaningful verification: **distance in km**!

Distance to Ice Edge

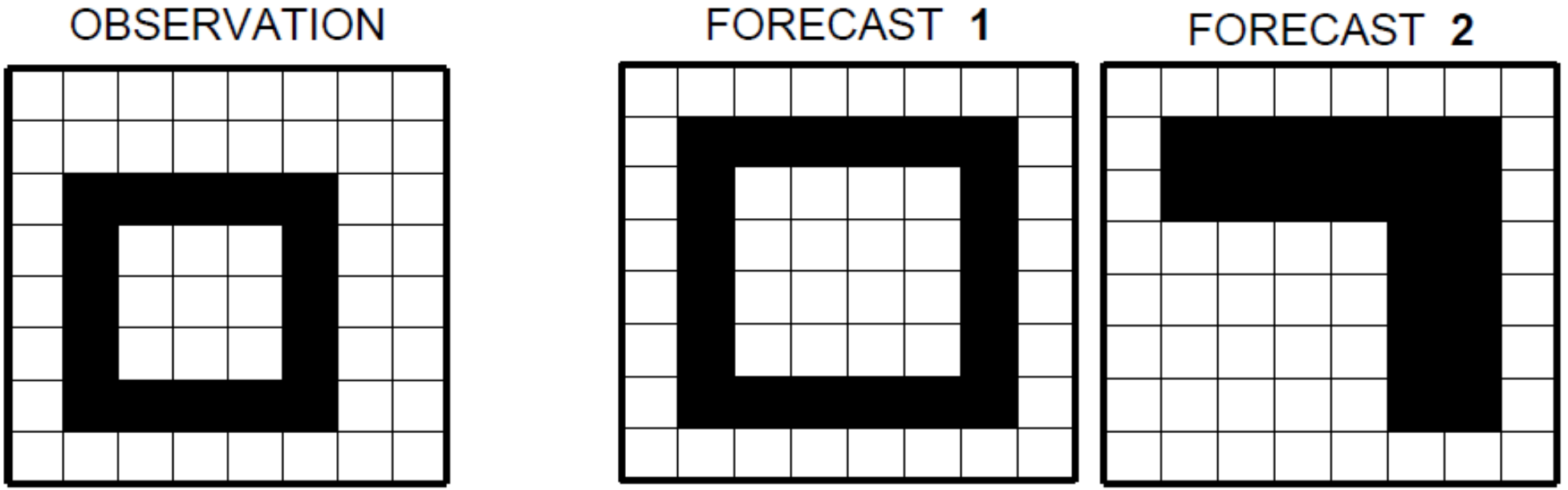
Courtesy of JF Lemieux (MRD-EC)

1. Thresholding: identify forecast and obs ice edges.
2. For each ice-edge pixel, evaluate the (min) distance between ice edges.
3. Consider mean, max (Hausdorff), median (Partial Hausdorff) distance.



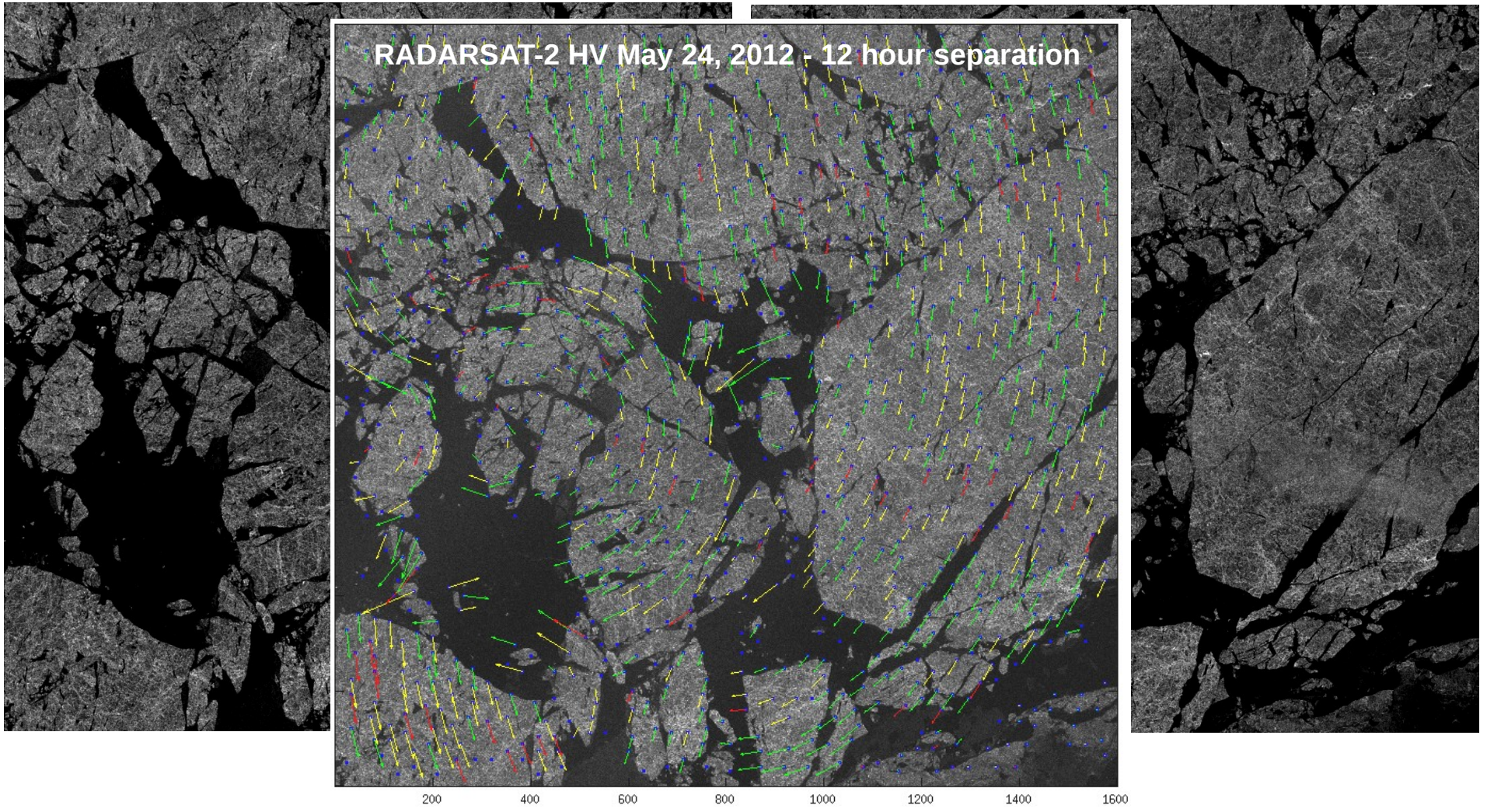
Verification of sea-ice extent: the Baddeley (1992) Delta metric

- Hausdorff metrics is overly-sensitive to spurious separate pixels
- Baddeley Delta metric: replace max with Lp norm (p=2)



Δ= 0.5625	Δ= 0.96875
hits = 9; false alarms = 11; misses = 7; corr.rej. = 37	

Ice drift trajectories

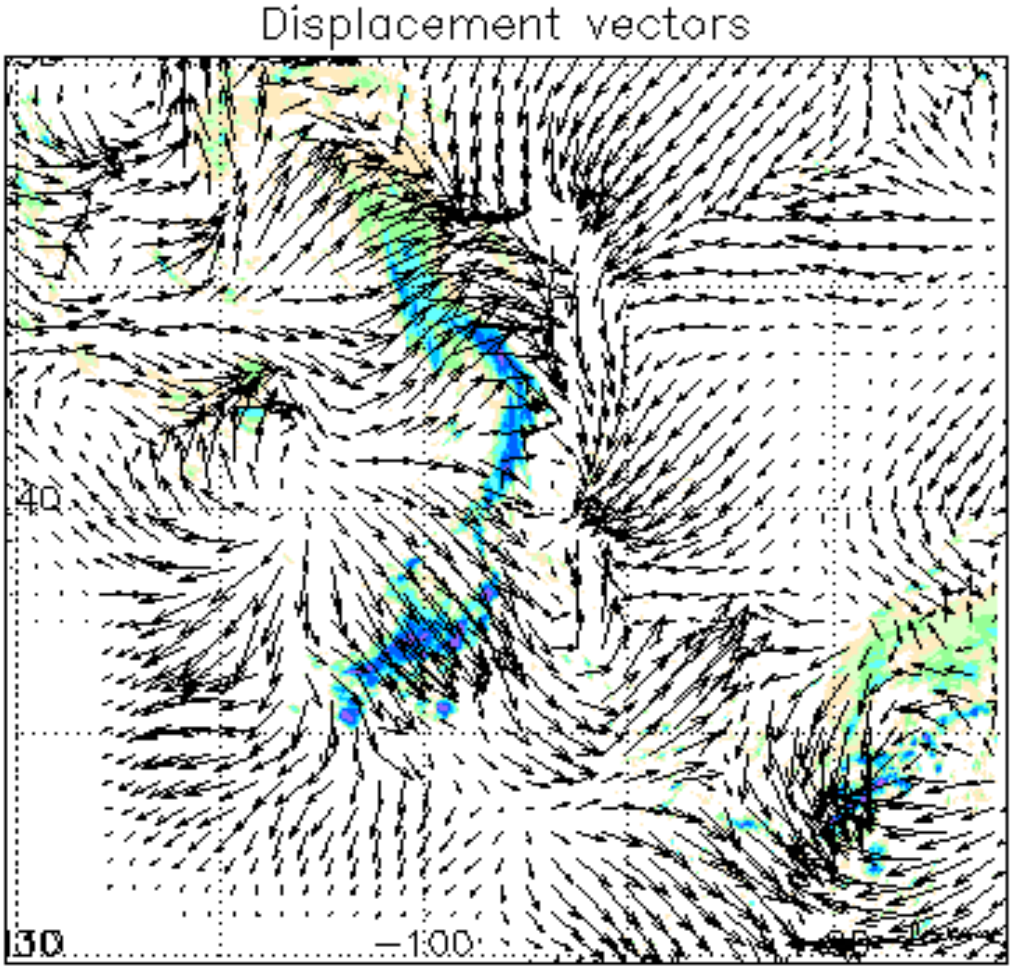


Automated sea ice tracking system: Komarov, A.S., Barber, D.G., (2014). Sea ice motion tracking from sequential dual-polarization RADARSAT-2 images. IEEE Transactions on Geoscience and Remote Sensing, vol. 52, no.1, pp.121-136.

Verify sea-ice motion

The automated sea-ice traking system produce a vector field similar to that produce by a **field-deformation (morphing) verification approaches:** investigate the possibility of verifying ice motion with a morphing techniques?

Alternative: use the **S1 score** to verify modelled versus obs sea-ice motion fields.



From Keil and Craig (2007, 2009)

The S1 Score

The S1 score was historically designed to assess the accurancy of the forecast in reproducing the gradients of pressure or geopotential height, in consideration of the relationship of these gradients to the wind field.

Dy = forecast pressure gradient between two adjacent grid-cells
Dx = obs pressure gradient between two adjacent grid-cells

$$S1 = \frac{\sum_{\text{adjacent pairs}} |Dy - Dx|}{\sum_{\text{adjacent pairs}} \max\{|Dy|,|Dx|\}}$$