Environment and Climate Change Canada
Environnement et Changement climatique Canada

Canada

Sea-ice verification by using binary image distance metrics

B. Casati, JF. Lemieux, G. Smith, P. Pestieau, A. Cheng

<u>Talk outline</u>: the quest for an informative metric (from Hausdorff to Baddeley, and beyond …)

Variable: RIPS vs IMS sea-ice extent (sea-ice concentration > 0.5)
<u>Goal: analyze the metric behaviour.</u> Once it is understood how the metric responds to different types or errors, then we can perform the verification of operational products ...

Context: Existing Verification Techniques

Traditional (point-by-point) methods:

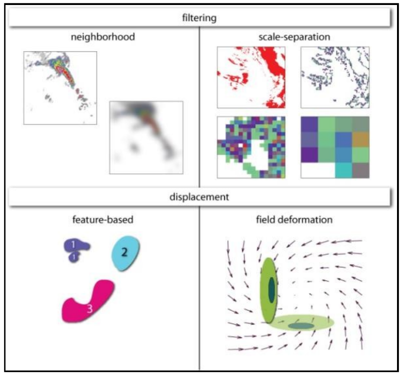
1. graphical summary (scatter-plot, box-plot, quantile-plots).
2. Continuous scores (MSE, correlation).
3. Categorical scores from contingency tables (FBI,HSS,PC).
4. Probabilistic verification (Brier, CRPS, rank histogram, reliability diagram).

Extreme dependency scores: Ferro and Stephenson 2011 (EDI,SEDI)

<u>There is no single technique</u>
which fully describes the complex observation-forecast relationship!
<u>Key factors</u>: verification end-user and purpose; (statistical) characteristics of the variable & forecast; available obs.

Spatial verification methods:

1. Scale-separation
2. Neighbourhood
3. Field-deformation
4. Feature-based



5. **Distance metrics for binary images**

- account for the **coherent spatial structure** (i.e. the intrinsic correlation between near-by grid-points) and the presence of **features**
- assess **location and timing errors** (separate from **intensity error**) in physical terms (e.g. km) – informative and meaningful verification
- account for **small time-space uncertainties** (avoid **double-penalty**)

Distance measures for binary images

Precipitation:

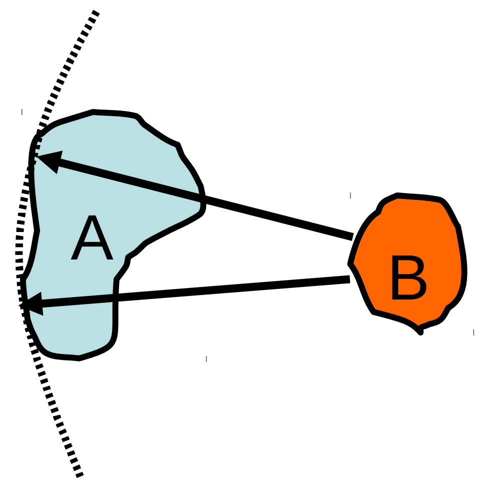
Gilleland et al.(2008), MWR 136
Gilleland et al (2011), W&F 26
Schwedler & Baldwin (2011), W&F 26
Venugopal et al. (2005), JGR-A 110
Zhu et al (2010), Atmos Res 102
Aghakouchak et al (2011), J.HydroMet 12
Brunet and Sills (2015), IEEE SPS 12

Sea-ice:

Heinrichs et al (2006), IEEE trans. GSRS 44
Dukhovskoy et al (2015), JGR-O 120
Hebert et al (2015), JGR-O 120

- Average distance
- K-mean
- Fréchet distance
- **Hausdorff metric**
  - **Modified Hausdorff**
  - **Partial Hausdorff**
- **Baddeley metric**
- Pratts' figure of merit

- Account for distance between objects, similarity in shapes, ...
- Binary images: alternative metrics to be used along with traditional categorical scores



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

Do we want a metric?

Note: in maths, metric = distance (error measure, the smaller the better)

Definition: a metric M between two sets of pixels A and B satisfies:
1. Positivity: $M(A,B) \geq 0$
2. Separation: $M(A,B) = 0$ if and only if $A = B$
3. Symmetry: $M(A,B) = M(B,A)$
4. Triangle Inequality: $M(A,C) + M(C,B) \geq M(A,B)$

Metrics are mathematically sound! … but, are they useful?

The metrics' properties imply:
1. Measures the error (the smaller, the better).
2. Perfect score is achieved if <u>and only if</u> forecast = obs.
3. Result does not depend on order of comparison.
4. If M(O,F1) >> M(O,F2) it means that F2 is much better than F1, i.e. M(F1,F2) is significantly large (it separates forecasts according to their accuracy).

Hausdorff distance

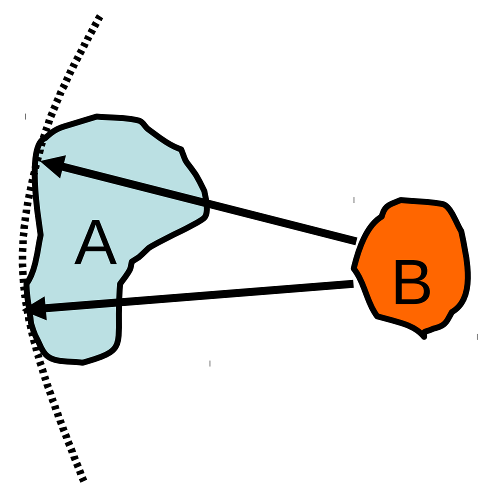
$$Haus(A,B) = max\{max_{a \in A}\, d(a,B)\, ;\, max_{b \in B}\, d(b,A)\}$$

The Hausdorff distance considers the *max* of the forward and backward distances:

$$d(A,B) = d(a,B)_{a \in A}$$

$$d(B,A) = d(b,A)_{b \in B}$$

Note: <u>backward and forward distances are not symmetric</u>: the “external” *max* enables symmetry!
The Hausdorff distance is a metric.

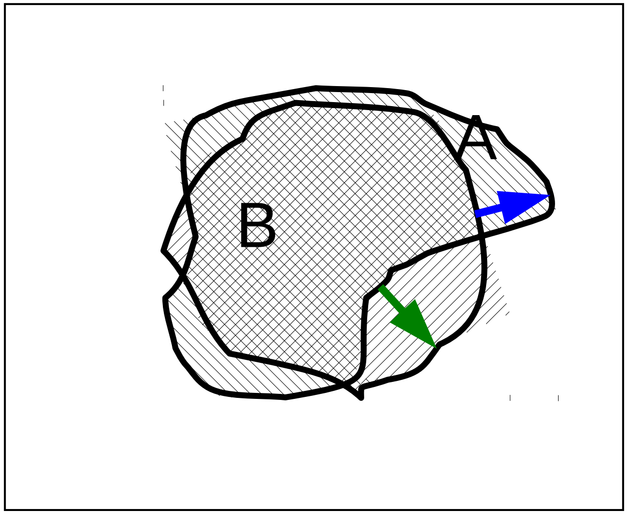


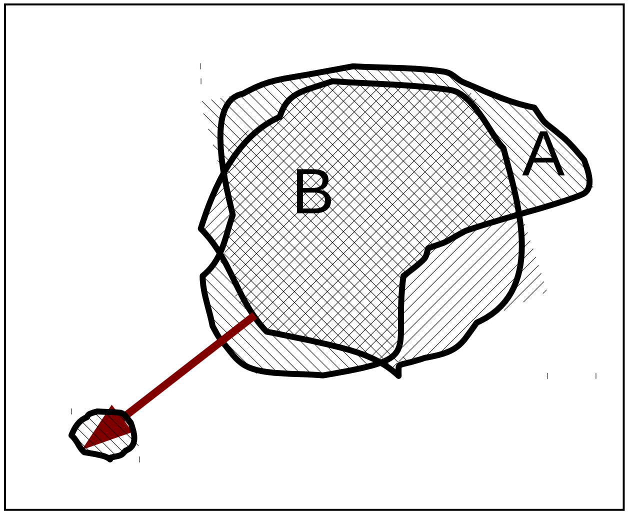
$$H(A,B) = \max\left\{\sup_{a \in A}\left(d(a,B)\right), \sup_{b \in B}\left(d(b,A)\right)\right\}$$

Hausdorff metric is sensitive to the distance between features

Shortcomings of the Hausdorff distance

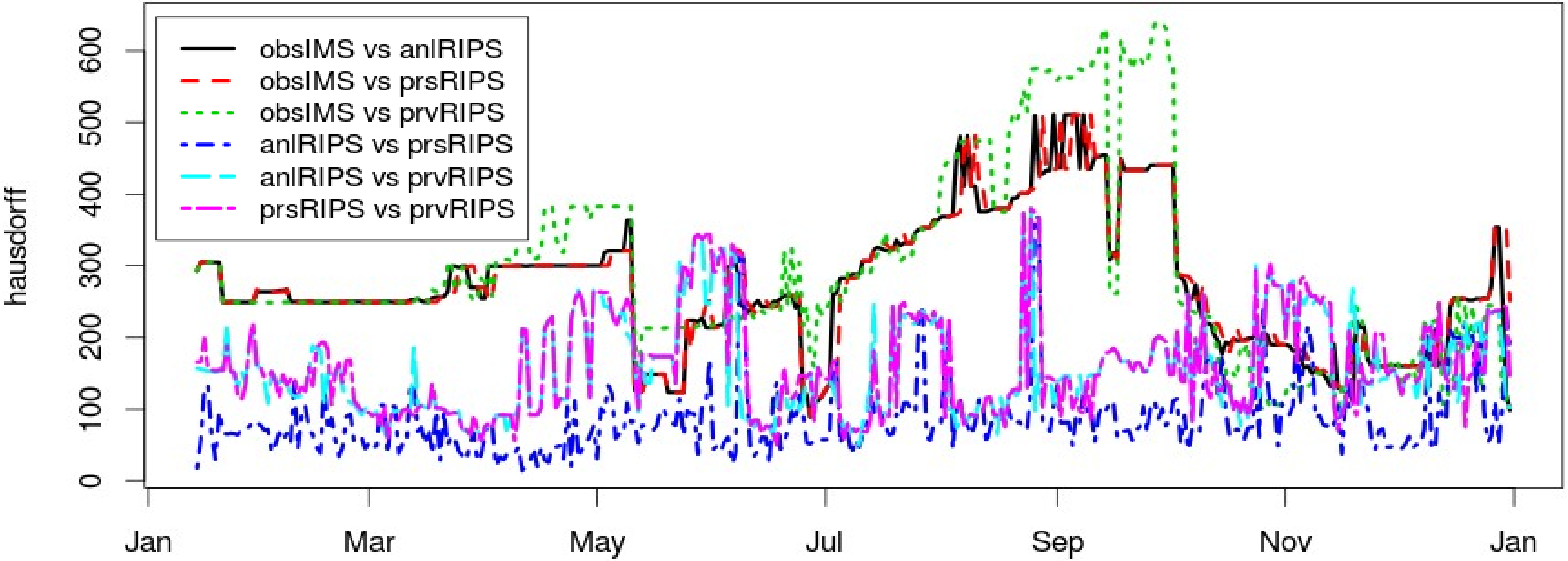
$$Haus(A,B) = max\{max_{a \in A}\, d(a,B)\,;\, max_{b \in B}\, d(b,A)\}$$





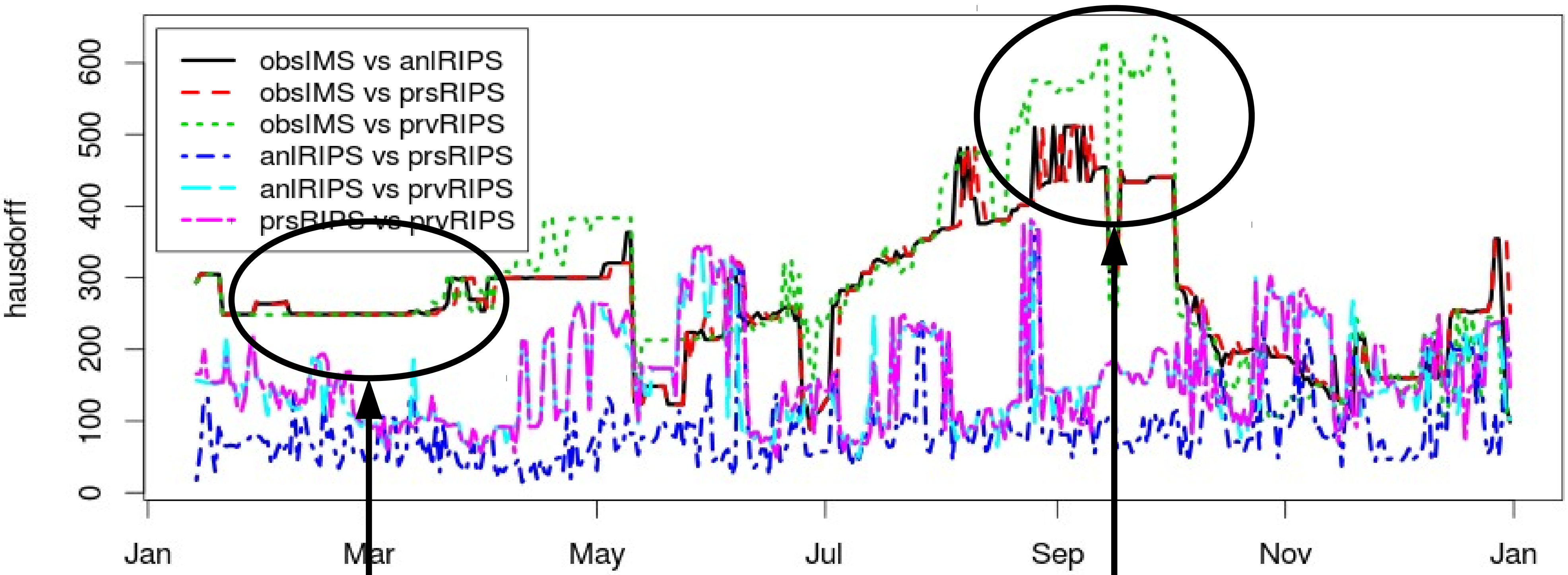
Because defined by using the *max*, the Hausdorff distance is overly sensitive to noise and **outliers**!
Example: spurious separated pixels associated with land-fast ice, which are generated by the RIPS forecast but are not visible in satellite products, lead to overly pessimistic / misleading scores.

Hausdorff distance, RIPS vs IMS



- Verification within RIPS products (bottom 3 lines) lead to better (smaller) scores than verification of RIPS products versus IMS obs (top 3 lines).
- RIPS analysis behaves as RIPS persistence (pers = 48h lag analysis)
- We focus on IMS obs versus RIPS forecast and IMS obs versus RIPS analysis: correlated behaviour → differences between RIPS forecast and IMS obs is directly inherited from RIPS analysis

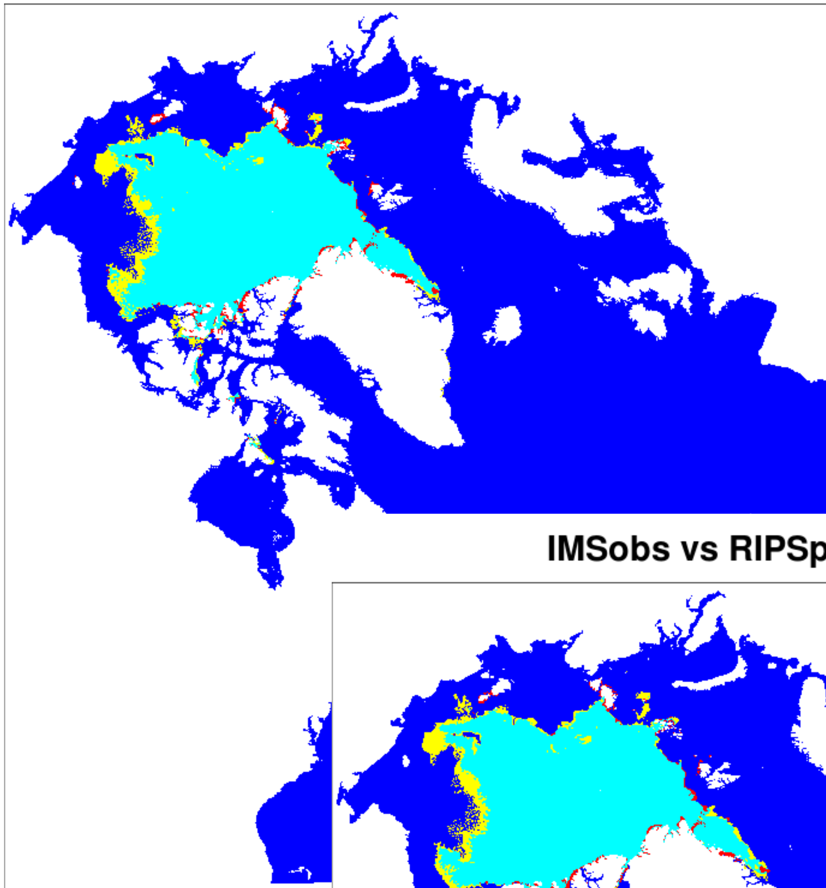
Hausdorff distance, RIPS vs IMS



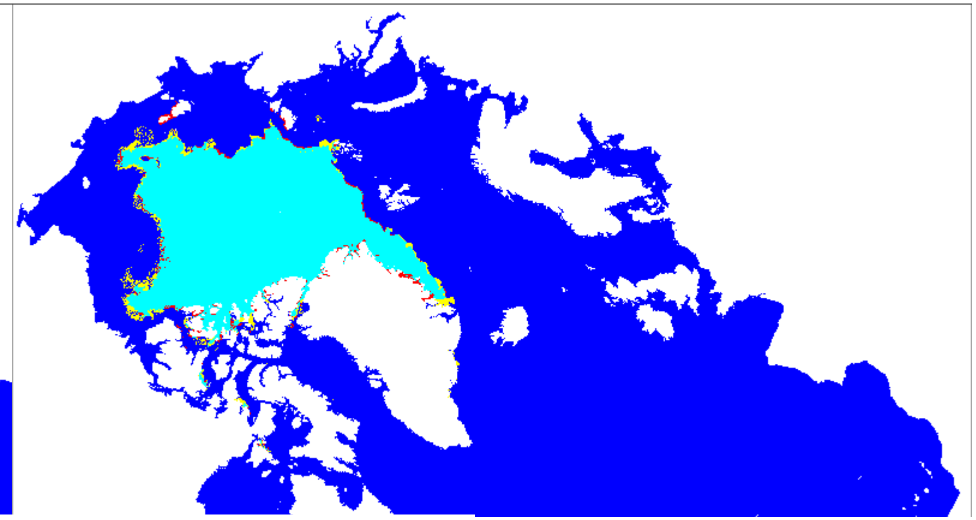
1st and 20th Sept: large error, prv > anl
Instead 20 better than 1, prv~anl

20/2,1/3,10/3:
small constant error, prv = anl
Instead: 20/2 better, prv > anl

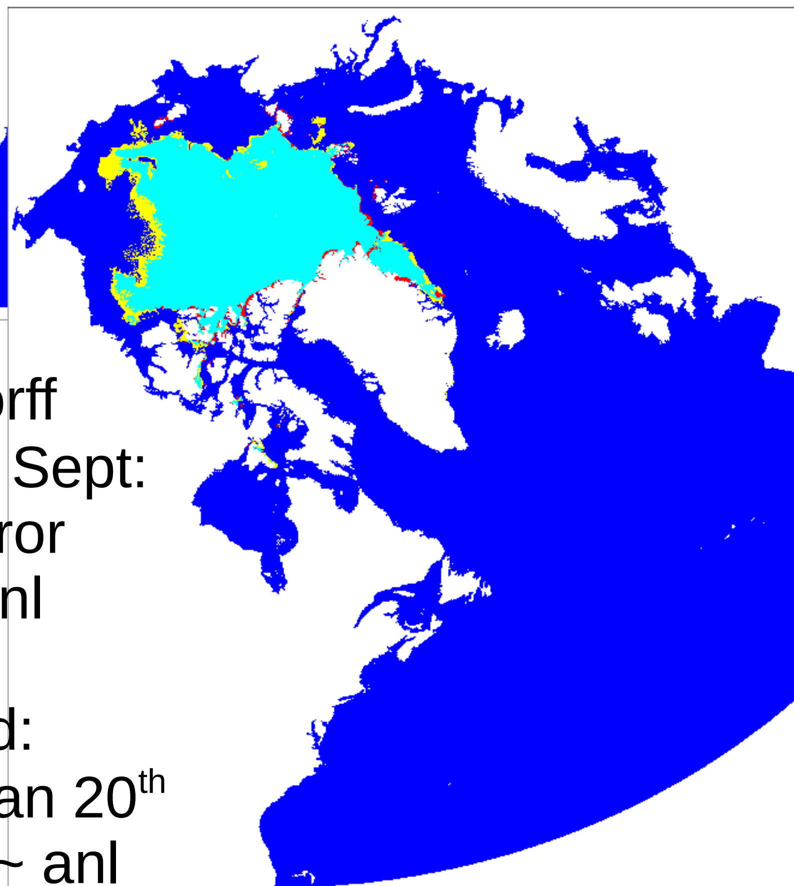
IMSobs vs RIPSanl on 20110901



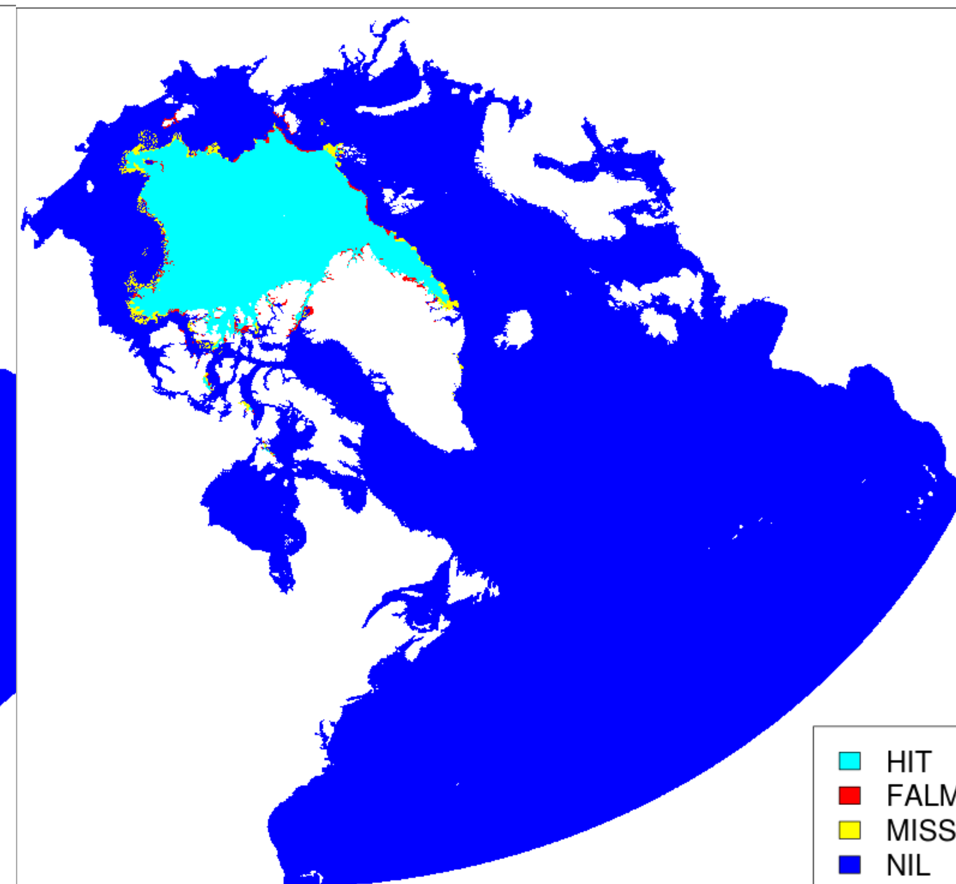
IMSobs vs RIPSanl on 20110920



IMSobs vs RIPSprv on 20110901



IMSobs vs RIPSprv on 20110920



Hausdorff
1^{st} and 20^{th} Sept:
large error
prv > anl

Instead:
1^{st} worse than 20^{th}
Sept, prv ~ anl

Partial and Modified Hausdorff Distances

$$PartHaus(A,B) = max\{q_{0.50}\, d(a,B)_{a \in A}\, ;\, q_{0.50}\, d(b,A)_{b \in B}\}$$

$$ModHaus(A,B) = max\{mean_{a \in A}\, d(a,B)\, ;\, mean_{b \in B}\, d(b,A)\}$$

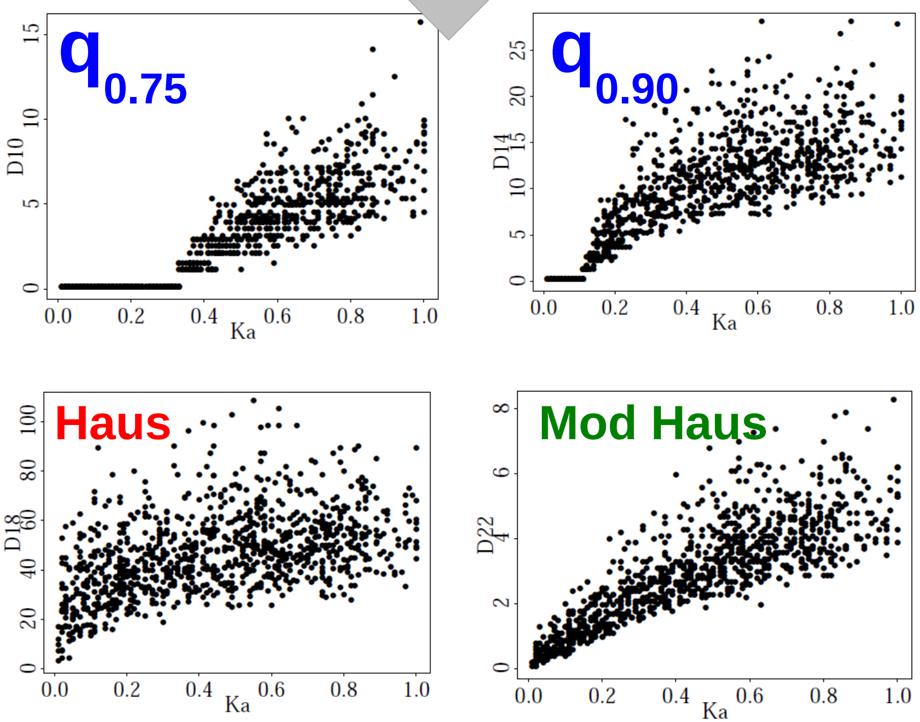
The partial / modified Hausdorff distances consider a quantile / the mean of the forward and backward distances.

Note:
The partial Hausdorff distance does not satisfy the separation property, nor the triangle inequality. The modified Hausdorff distance does not satisfy the triangle inequality:
<u>The partial and modified Hausdorff distances are not metrics</u>!

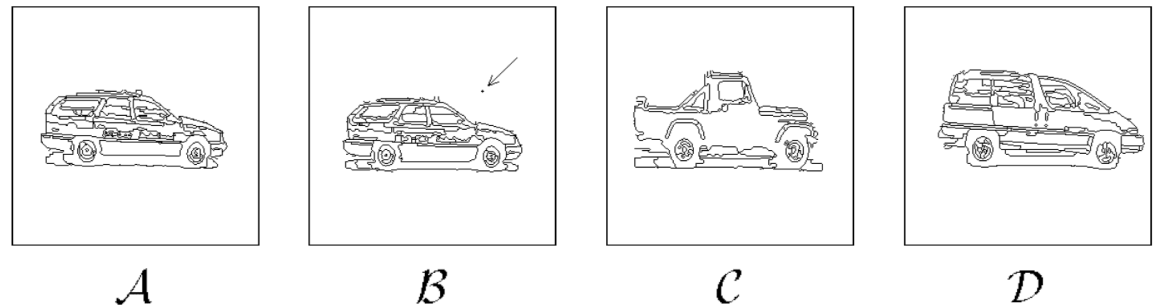
Dubuisson and Jain (1994) "A Modified Hausdorff Distance for Object Matching" Proc. International Conference on Pattern Recognition, Jerusalem (Israel) page 566-568.

Test sensitivity to noise:

- Hausdorff is overly-sensitive
- PartHaus does not separate
- ModHaus desired response



Test distances for edge detection



D_{10}				
	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$	$\mathcal{D}$
$\mathcal{A}$	0	**2**	5	3
$\mathcal{B}$	**2**	0	7	2
$\mathcal{C}$	5	7	0	6
$\mathcal{D}$	3	2	6	0

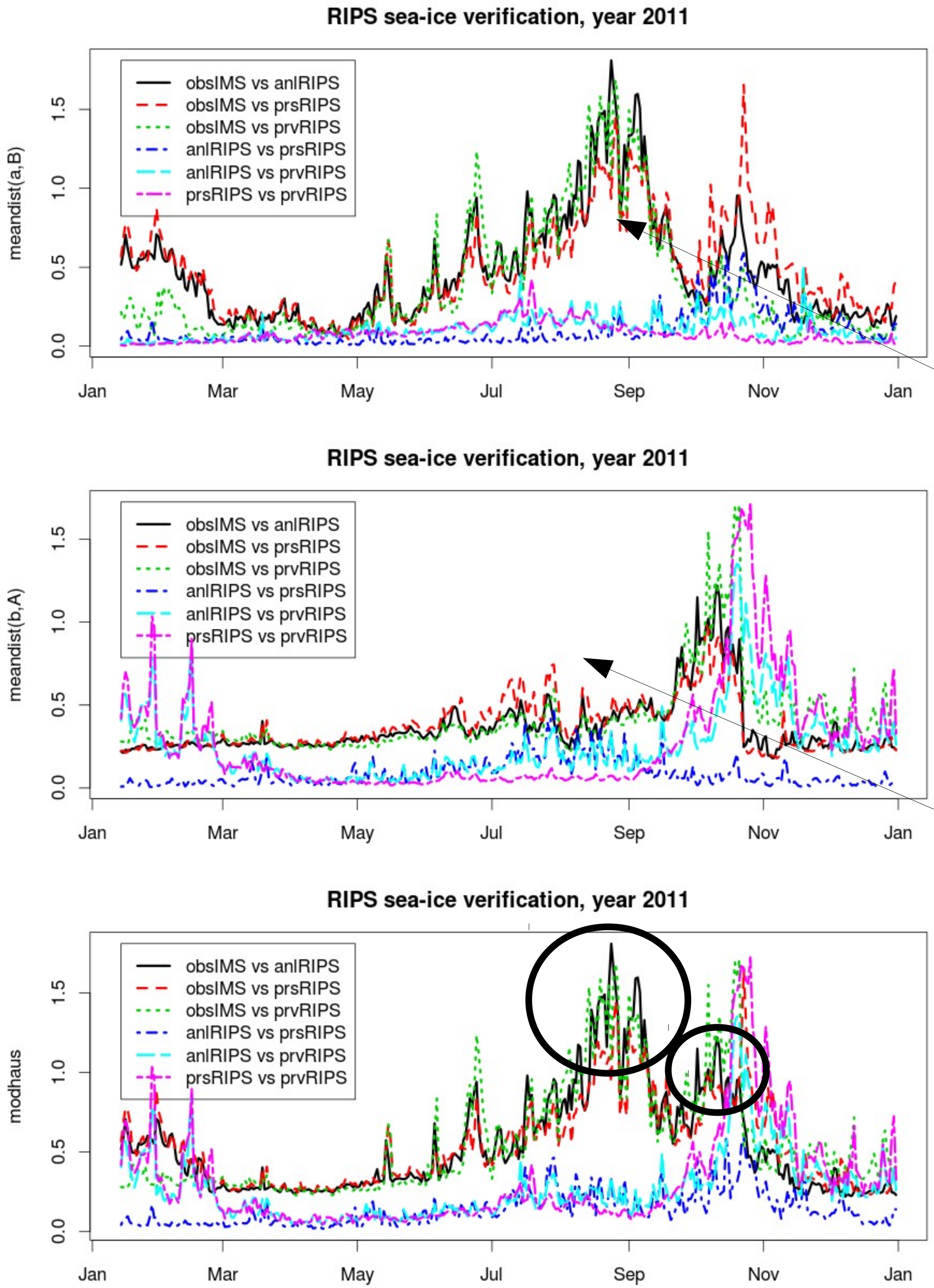
D_{14}				
	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$	$\mathcal{D}$
$\mathcal{A}$	0	**3**	10	6
$\mathcal{B}$	**3**	0	10	7
$\mathcal{C}$	10	10	0	12
$\mathcal{D}$	6	7	12	0

D_{18}				
	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$	$\mathcal{D}$
$\mathcal{A}$	0	**32**	22	32
$\mathcal{B}$	**32**	0	107	25
$\mathcal{C}$	22	107	0	36
$\mathcal{D}$	32	25	36	0

D_{22} (MHD)				
	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$	$\mathcal{D}$
$\mathcal{A}$	0	**1**	6	4
$\mathcal{B}$	**1**	0	6	4
$\mathcal{C}$	6	6	0	6
$\mathcal{D}$	4	4	6	0

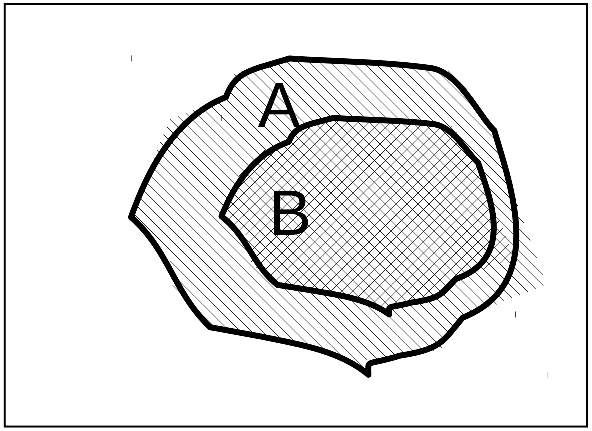
Table 2: D_{10}, D_{14}, D_{18}, and D_{22} for the four objects shown in Figure 4. Note that MHD has the best discriminatory power for object matching.

Modified Hausdorff, RIPS vs IMS



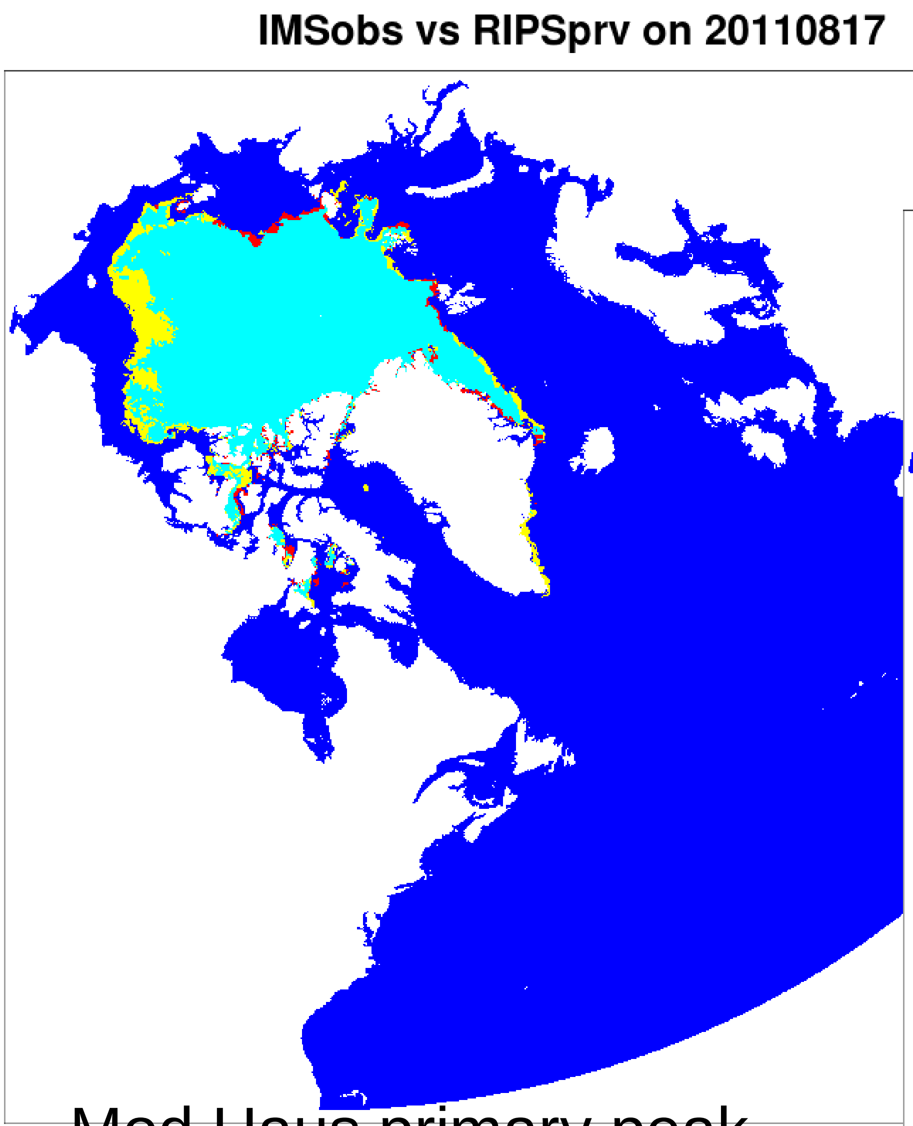
Reminder: the backward and forward (mean) distances are not symmetric:

$d(A,B) = d(a,B)_{a \in A} \neq 0$

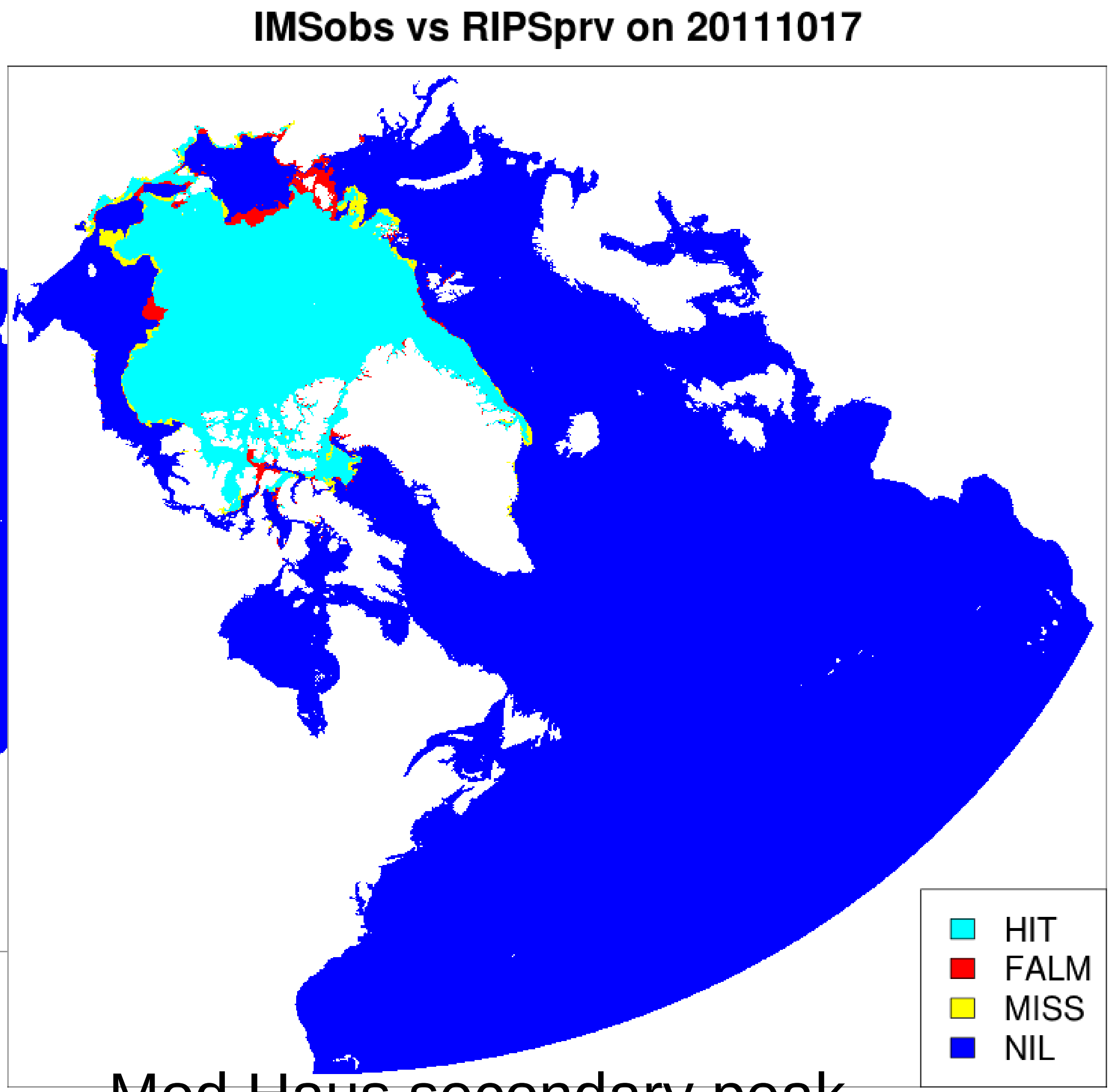


$d(B,A) = d(b,A)_{b \in B} = 0$

Differences are due to inclusion of sea-ice features, sea-ice extent over and underestimation.
Asymmetry is informative!



Mod Haus primary peak,
fwd >> bkw:
RIPS forecast / analysis
underestimate the sea-ice extent because melt ponds are assimilated as water



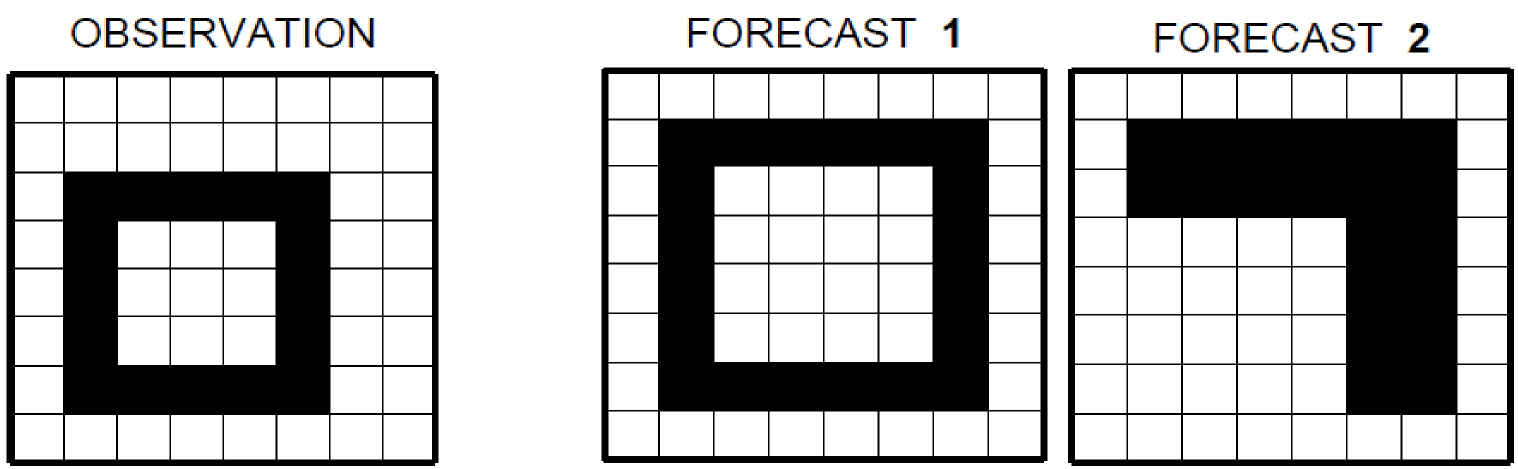
Mod Haus secondary peak,
bkw >> fwd:
RIPS forecast overestimates the sea-ice extent

The Baddeley (1992) Delta (Δ) metric

$$Haus(A,B) = max\{max_{a \in A}\, d(a,B)\,;max_{b \in B}\, d(b,A)\} =$$

$$= max_{x \in X}|d(x,A) - d(x,B)| = L_\infty$$

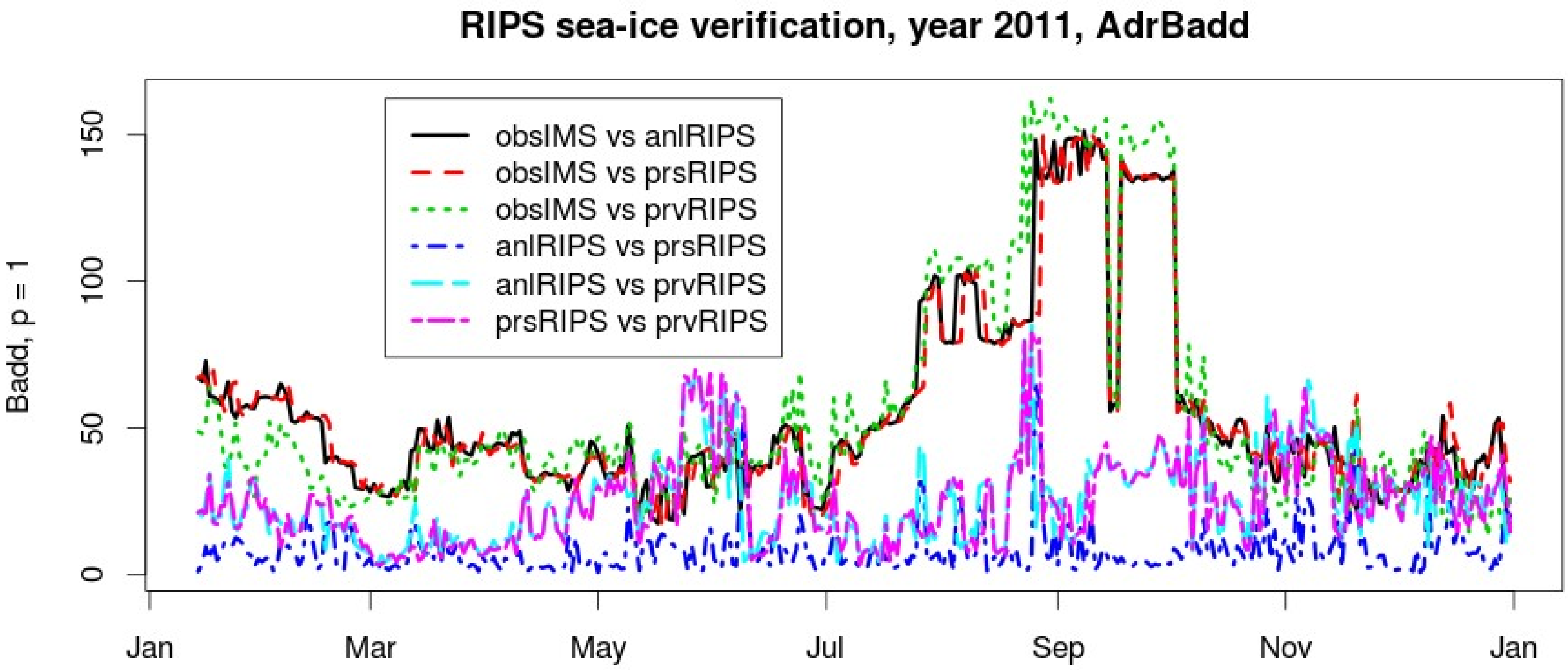
$$Badd(A,B) = \sqrt[p]{mean_{x \in X}|d(x,A) - d(x,B)\,|^p} = L_p \qquad p = 1,2,\ldots$$



The Baddeley Delta accounts for the similarity in shape

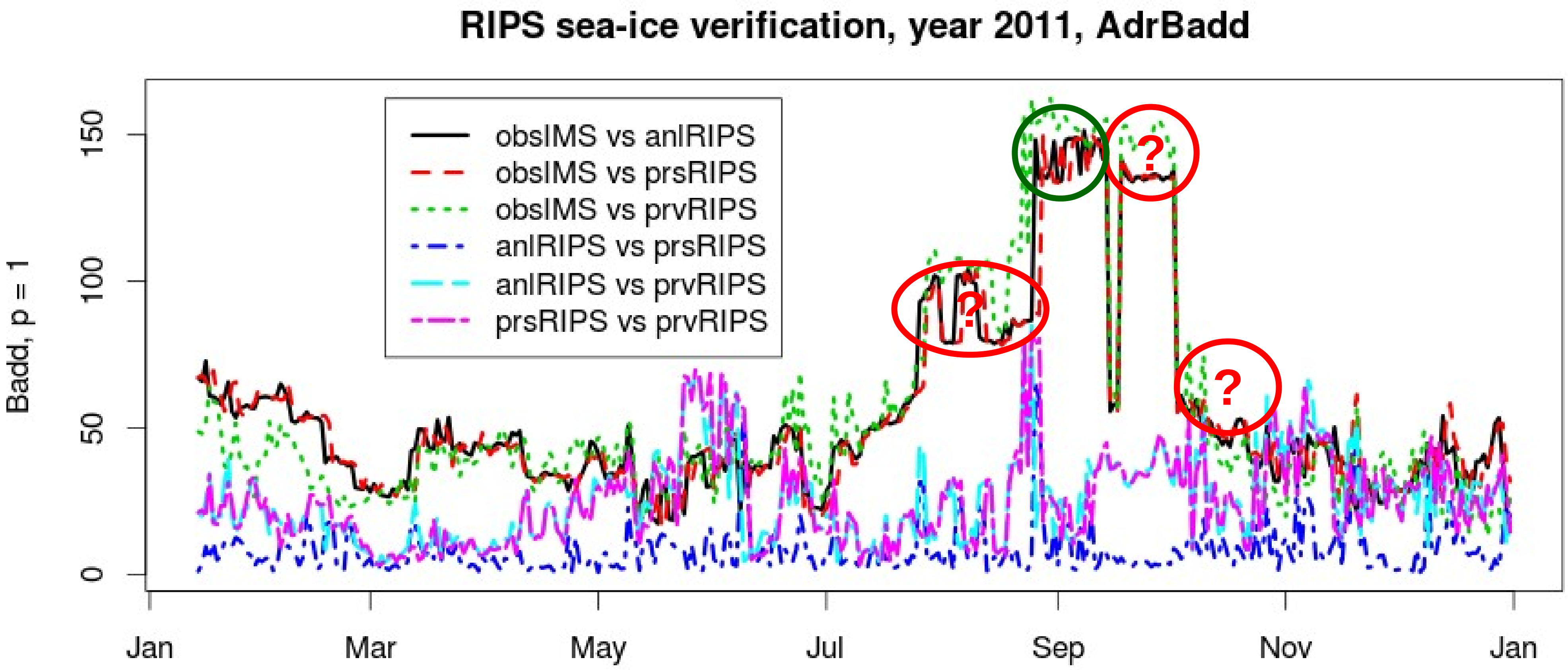
Δ= 0.5625	Δ= 0.96875
hits = 9; false alarms = 11; misses = 7; nils = 37	

Baddeley Delta Metric, RIPS vs IMS



The Baddeley metric behaves similarly to the Hausdorff distance: poor discriminatory power!!

Baddeley Delta Metric, RIPS vs IMS



The Baddeley metric behaves similarly to the Hausdorff distance: poor discriminatory power!!

- Large misses in late August, early September
- Large false alarms in mid October
- 20th September better than 1st September

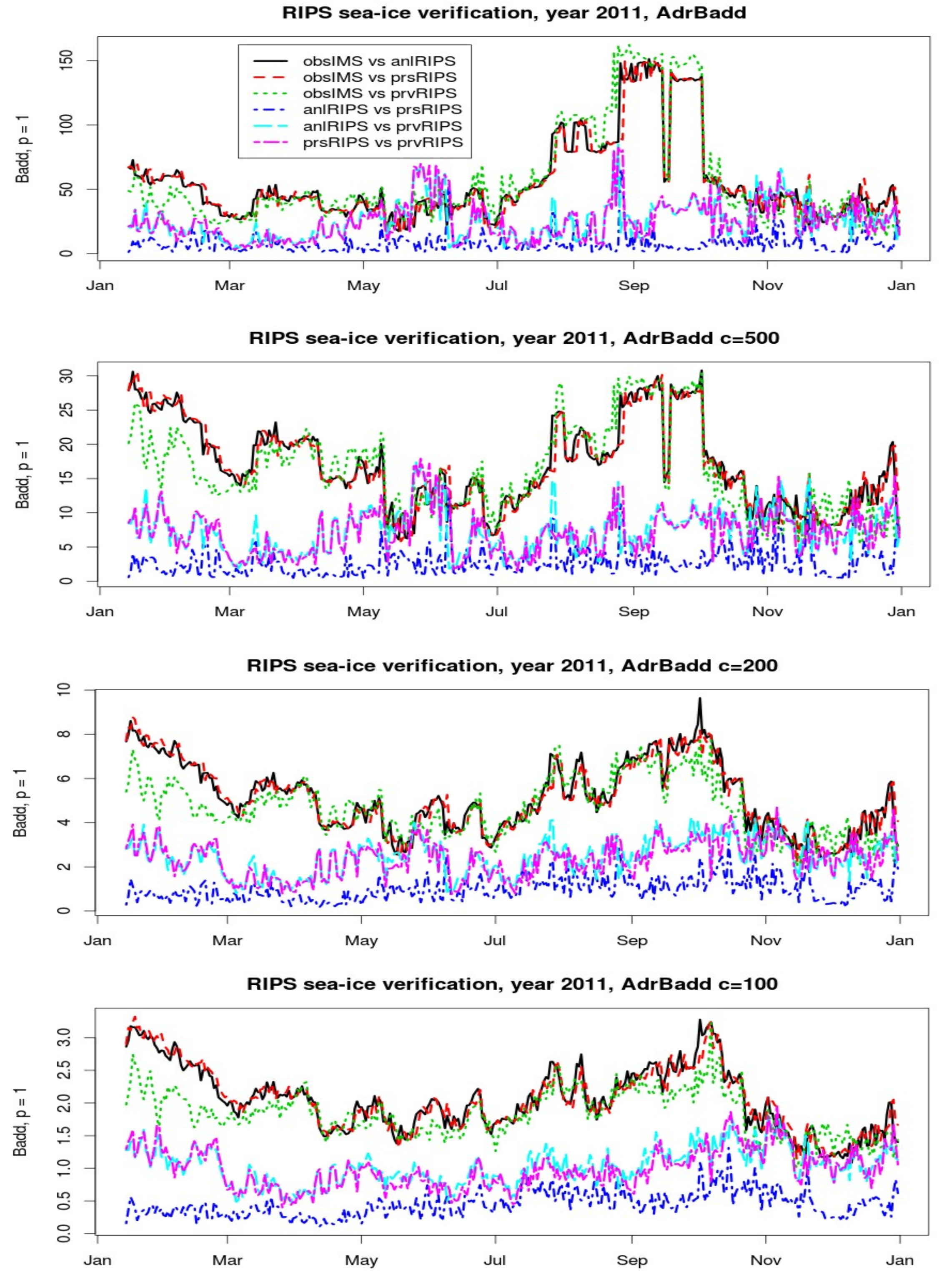
Shortcomings of the Baddeley Δ metric

The Baddeley metric is sensitive to the domain size: addition of zeros increases the distance!

$$Badd(A,B)=\left(\mathrm{mean}_{x\epsilon X}|d(x,A)-d(x,B)|^{p}\right)^{1/p}$$

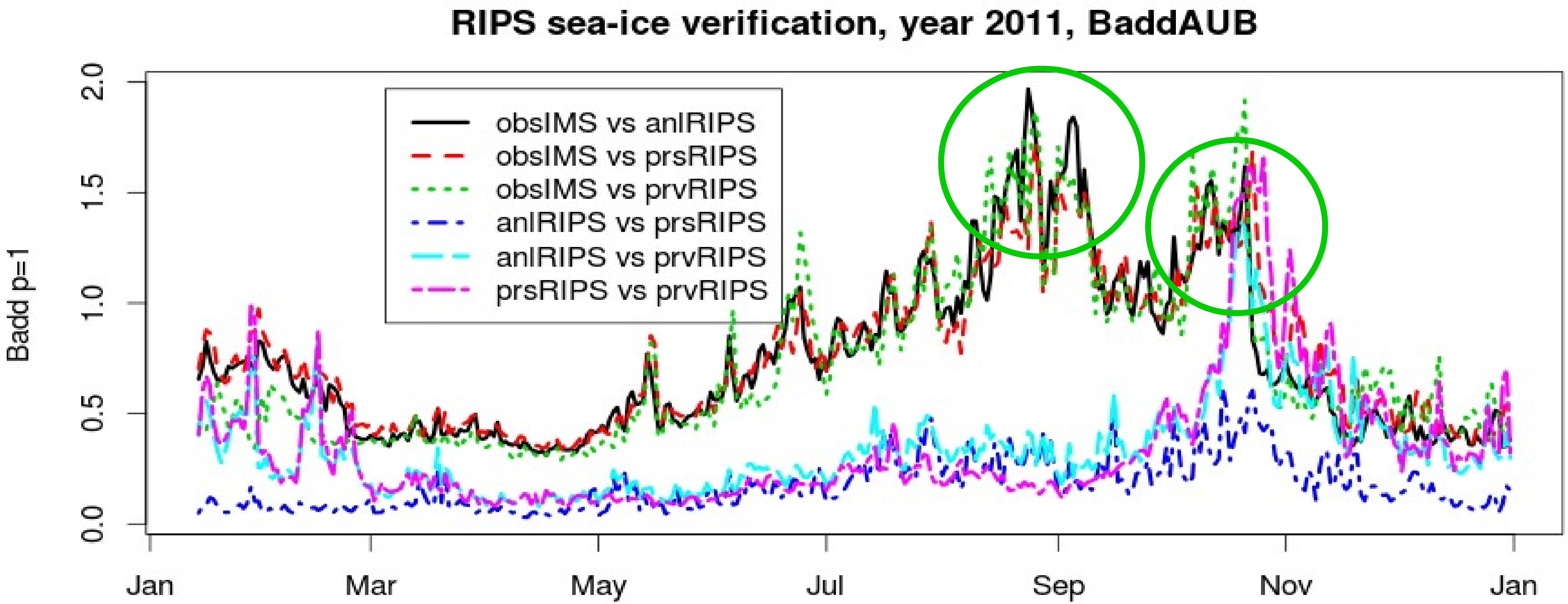
Solution 1:
C = cutoff distance
If d(x,A)>C, then d(x,A)=C
If d(x,B)>C, then d(x,B)=C

Solution 2:
Evaluate the Baddeley metric over AUB rather than over the whole X.



Baddeley Δ metric evaluated on AUB

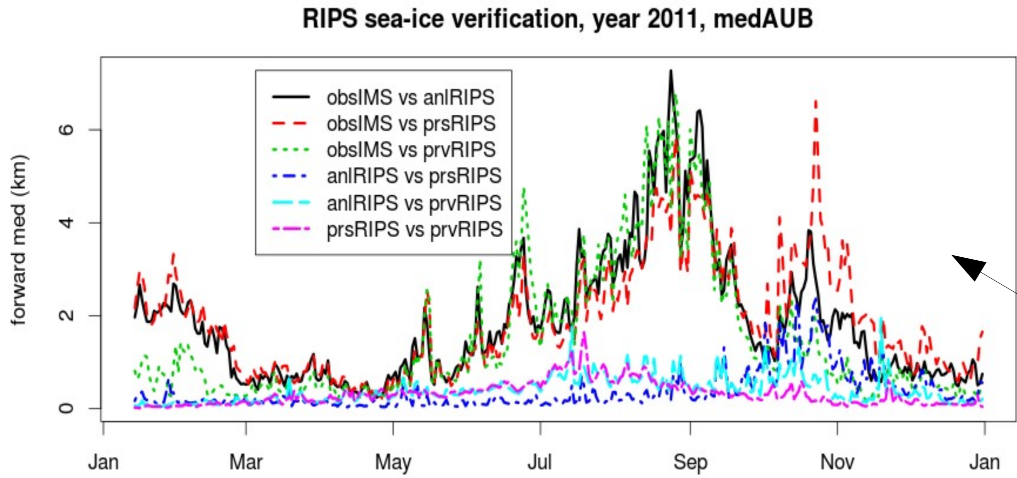
$$Badd_{AUB}(A,B)=\left[\frac{1}{n_{AUB}}\left(\sum_{a\in A\setminus B} d(a,B)^p+\sum_{b\in B\setminus A} d(b,A)^p\right)\right]^{1/p}$$



The Baddeley metric evaluated on AUB is capable of discriminating poor vs better performance (20th September better than 1st September), and correctly diagnoses large misses in late August / early September and large false alarms in mid October: is BaddAUB a metric?

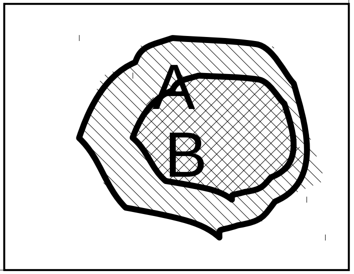
Distances in km

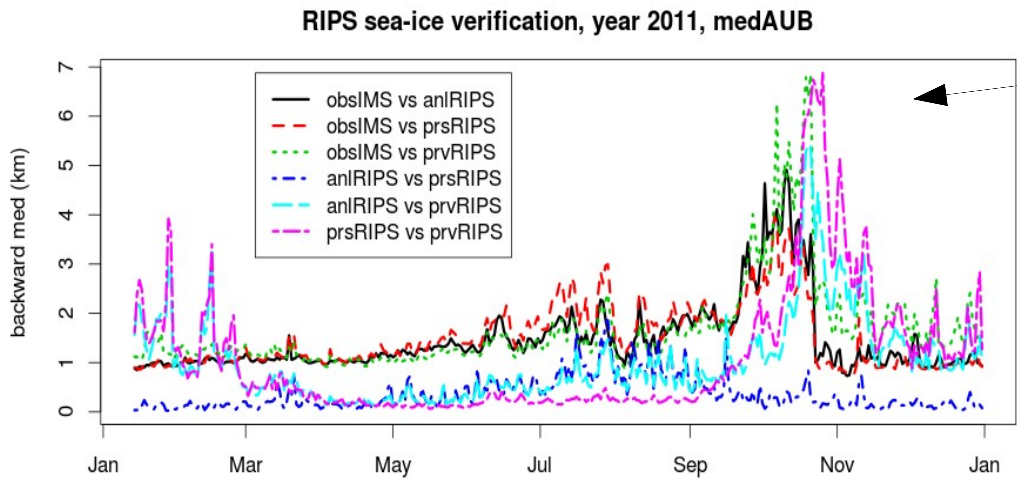
Technical but important detail: there is **no need to interpolate** forecast to obs grid!



Backward and forward mean distances (are not symmetric)

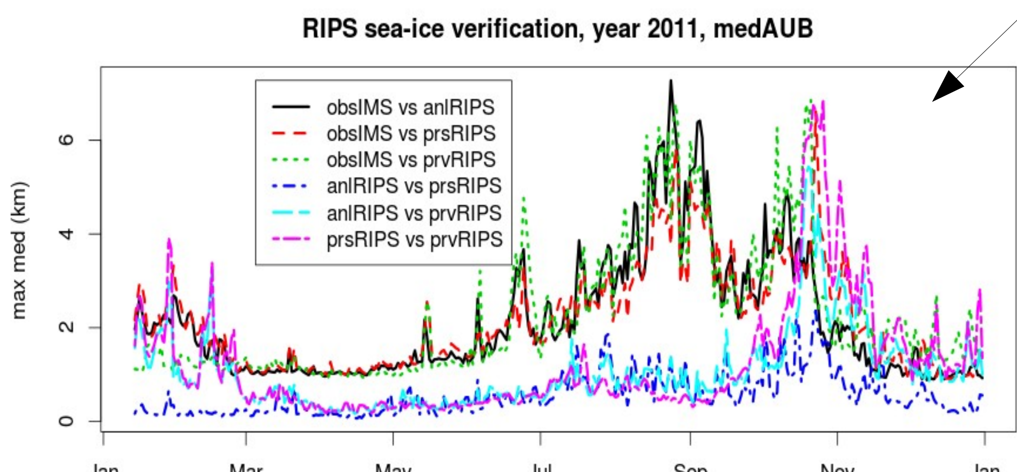
$$d(A,B) = d(a,B)_{a \in A} \neq 0$$
$$d(B,A) = d(b,A)_{b \in B} = 0$$

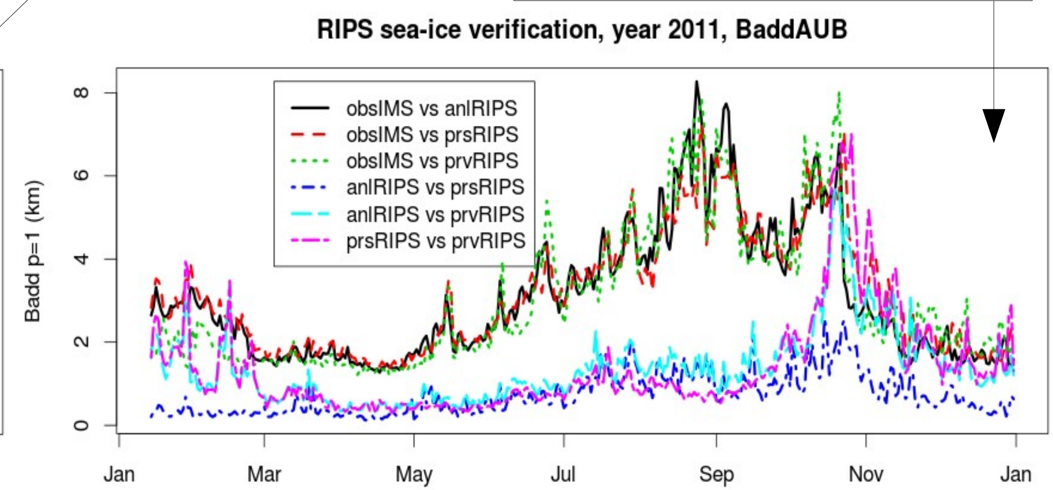




Modified Hausdorff

Baddeley metric evaluated on AUB





Conclusions and future work

Sea-ice verification by using the mean error distance, modified Hausdorff metric and Baddeley metric evaluated on AUB:

- agree with human perception / eye-ball verification
- is informative on false-alarms / misses,
- provides physical distances in km
- no interpolation needed

Hausdorff, Partial Hausdorff and Baddeley metric evaluated over the whole domain were found to be less informative and not robust.

Coming soon: apply the binary distance metrics to the ice-edge.
Sensitivity to edges present in IMS and not in RIPS: separate verification of Arctic Ocean vs Canadian channels ...

THANK YOU!

barbara.casati@canada.ca

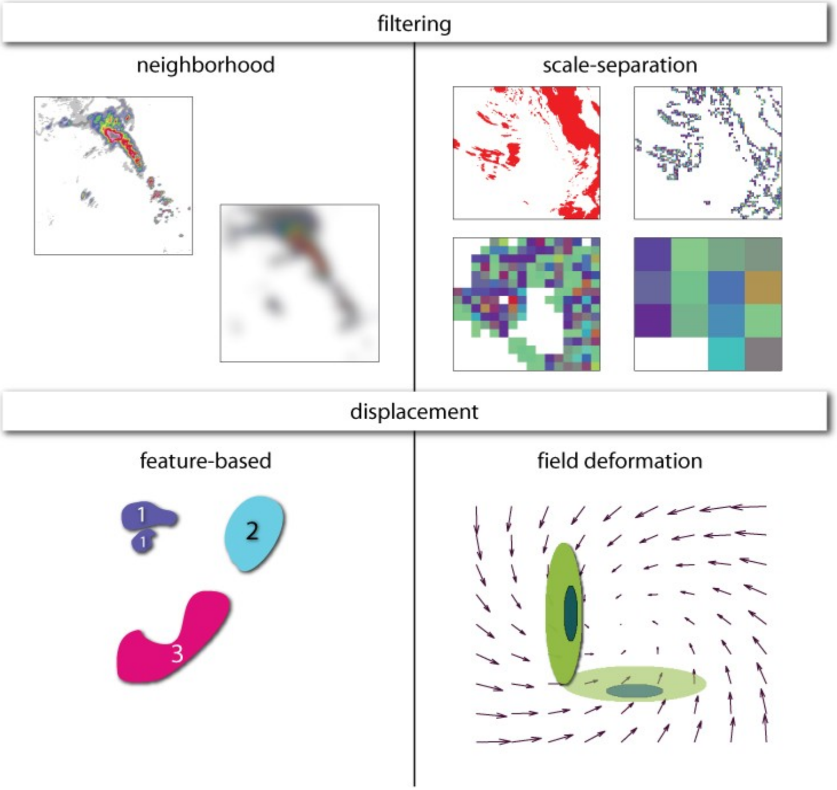
Verification Resources

http://www.cawcr.gov.au/projects/verification/

Forecast verification FAQ: web-page maintained by the WMO Joint Working Group on Forecast Verification Research (JWGFVR). Includes verification basic concepts, overview traditional and spatial verification approaches, links to other verification pages and verification software, key verification references.

http://www.ral.ucar.edu/projects/icp

Web page of the **Spatial Verification Inter-Comparison Project** (ICP), which now is entering its second phase (MesoVIC). Includes an *impressive list of references* for spatial verification studies.
Review article: Gilleland, E., D. Ahijevych, B.G. Brown, B. Casati, and E.E. Ebert, 2009: Intercomparison of Spatial Forecast Verification Methods. Wea. Forecasting, 24 (5), 1416 – 1430.



Thanks to Eric Gilleland
R package SpatialVx

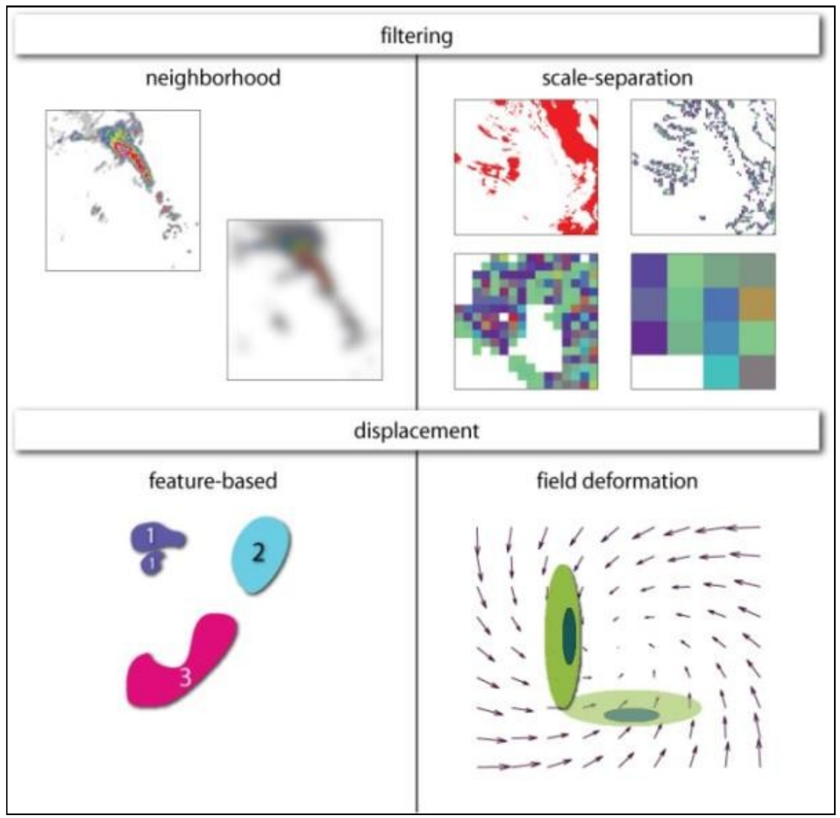
Extras 1
spatial verification approaches

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

Feature-based: evaluate attributes of isolated features



From Gilleland et al 2010

Scale-separation: analyse scale-dependency of forecast error

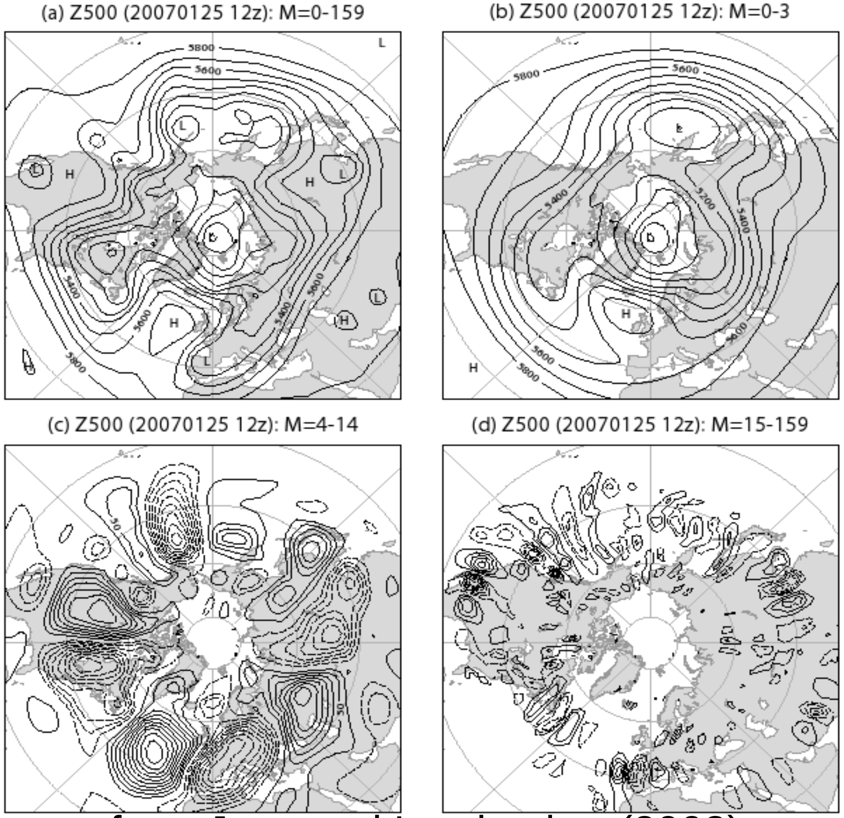
Field-deformation: use a vector and scalar field to morph forecast into obs

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

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)

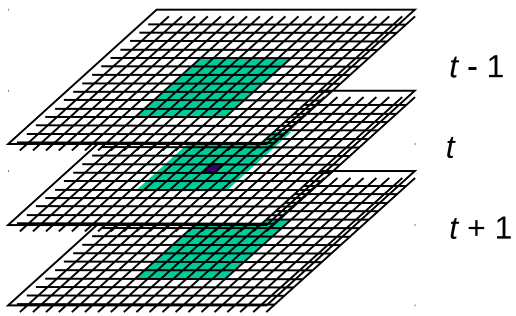


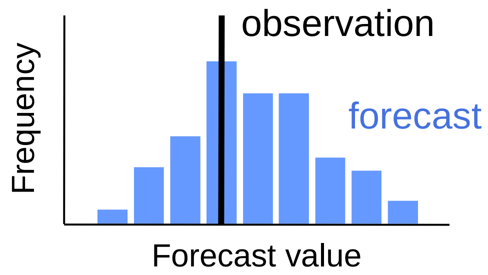
from Jung and Leutbecher (2008)

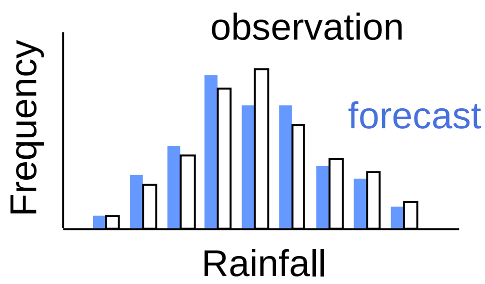
➔ Assess scale structure
➔ Bias, error and skill on different scales
➔ Scale dependency of forecast predictability (no-skill to skill transition scale)

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.

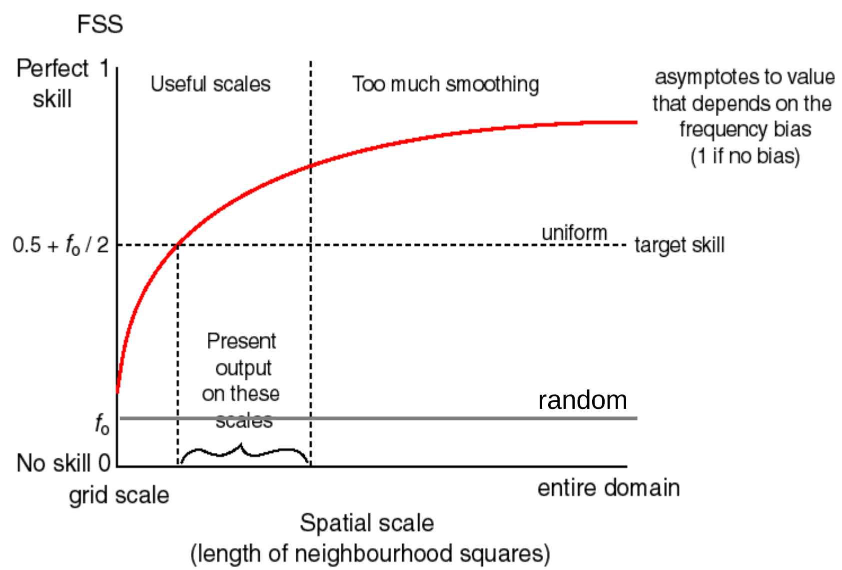






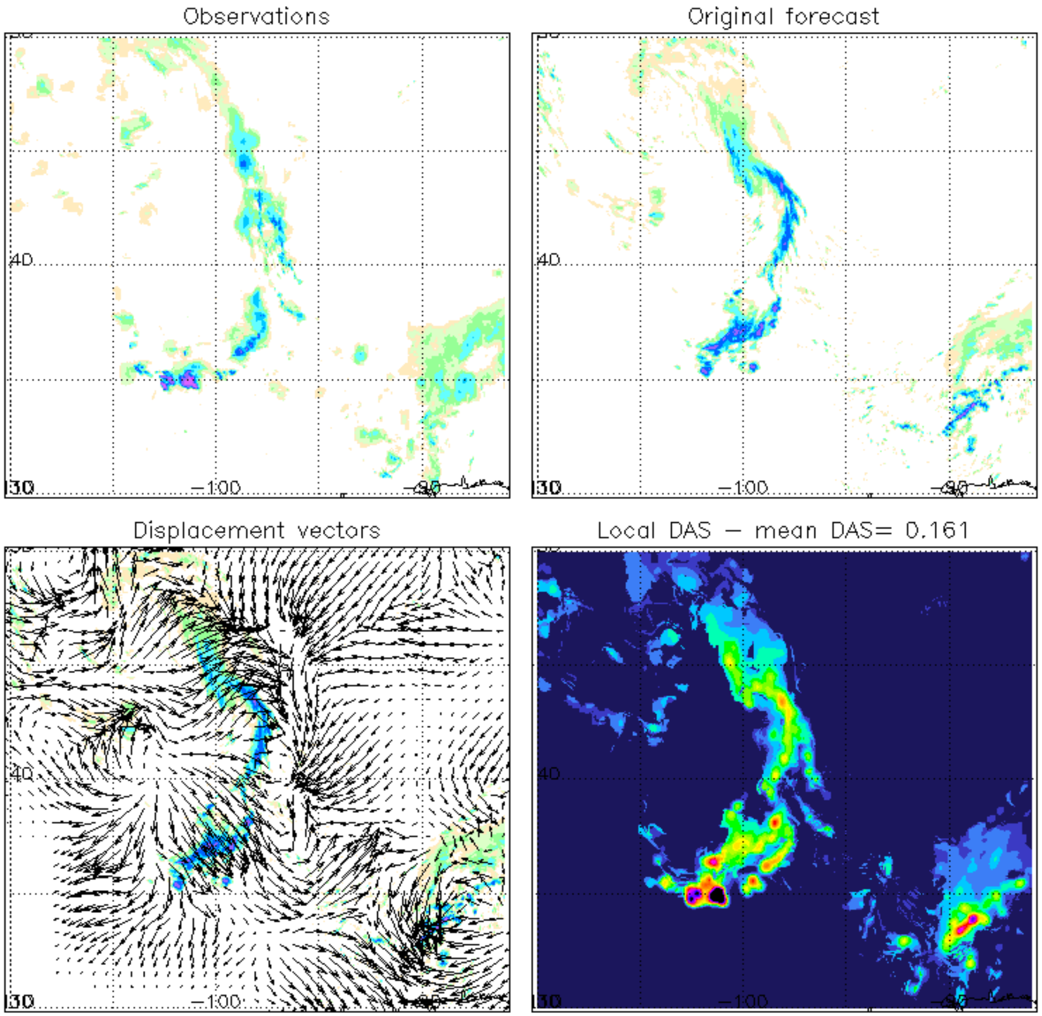
2. Perform verification over neighbourhoods of different sizes: verify deterministic forecast with probabilistic approach

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

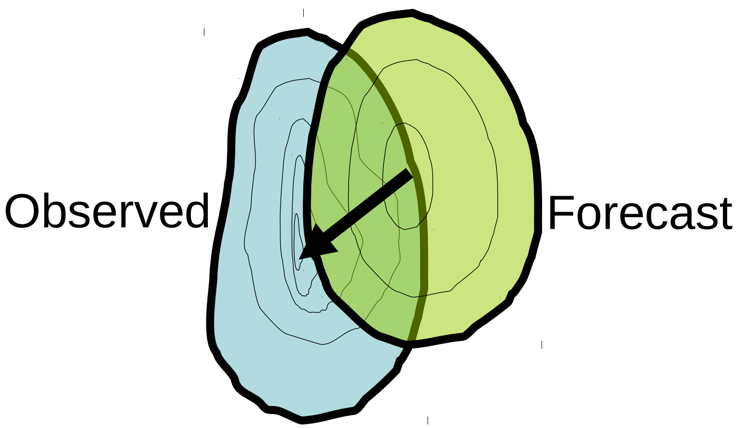


1. Use a vector (wind) field to deform the forecast field towards the obs field
2. Use an amplitude field to correct intensities of (deformed) forecast field to those of the obs field

- 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.

Extras 2
distance to ice-edge

Ice-edge verification

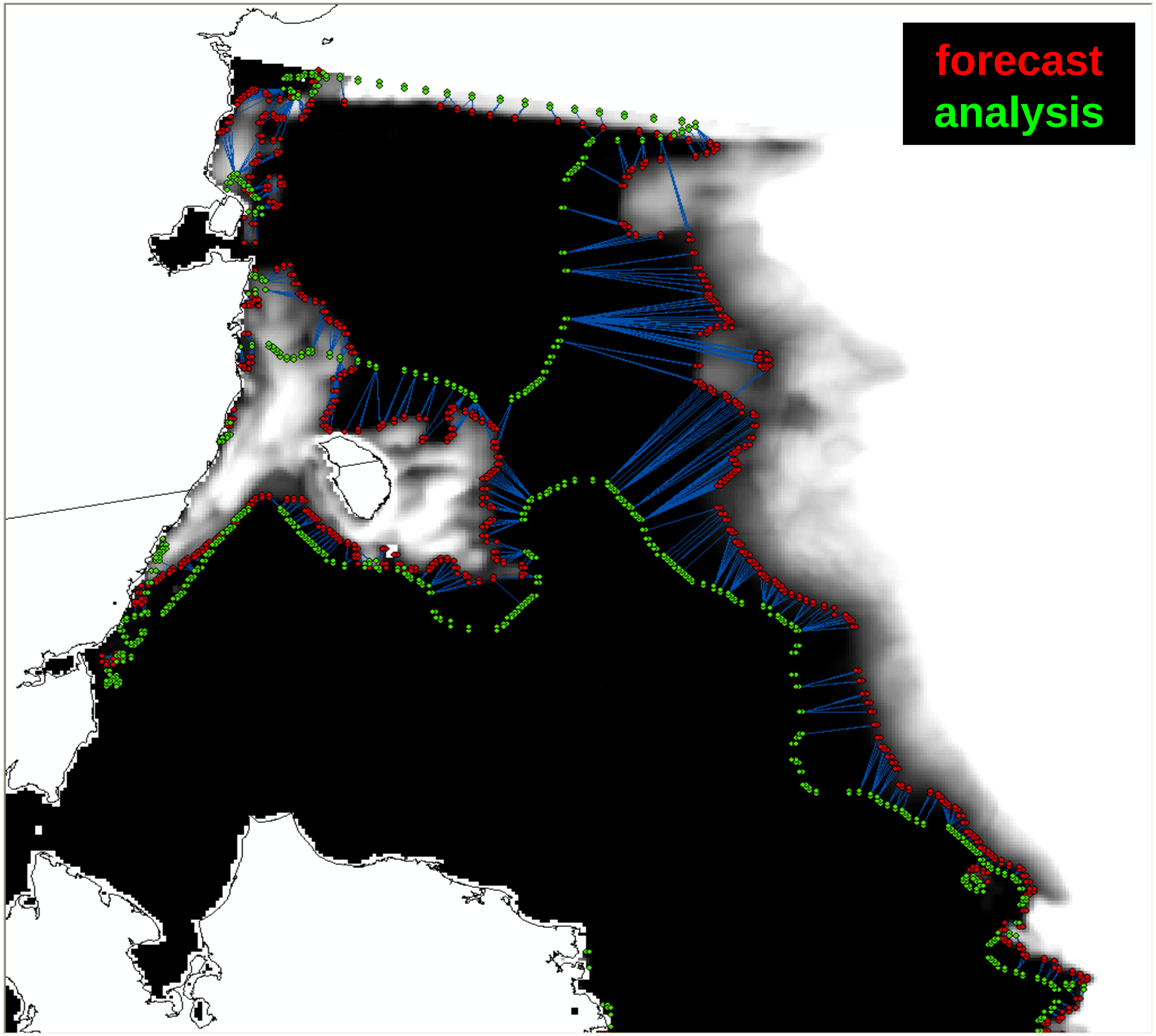


Image is courtesy of Angela Cheng (CIS)

Evaluate the distance between forecast and obs ice-edge by using the **Baddeley metric and** (partial and modified) **Hausdorff distances**

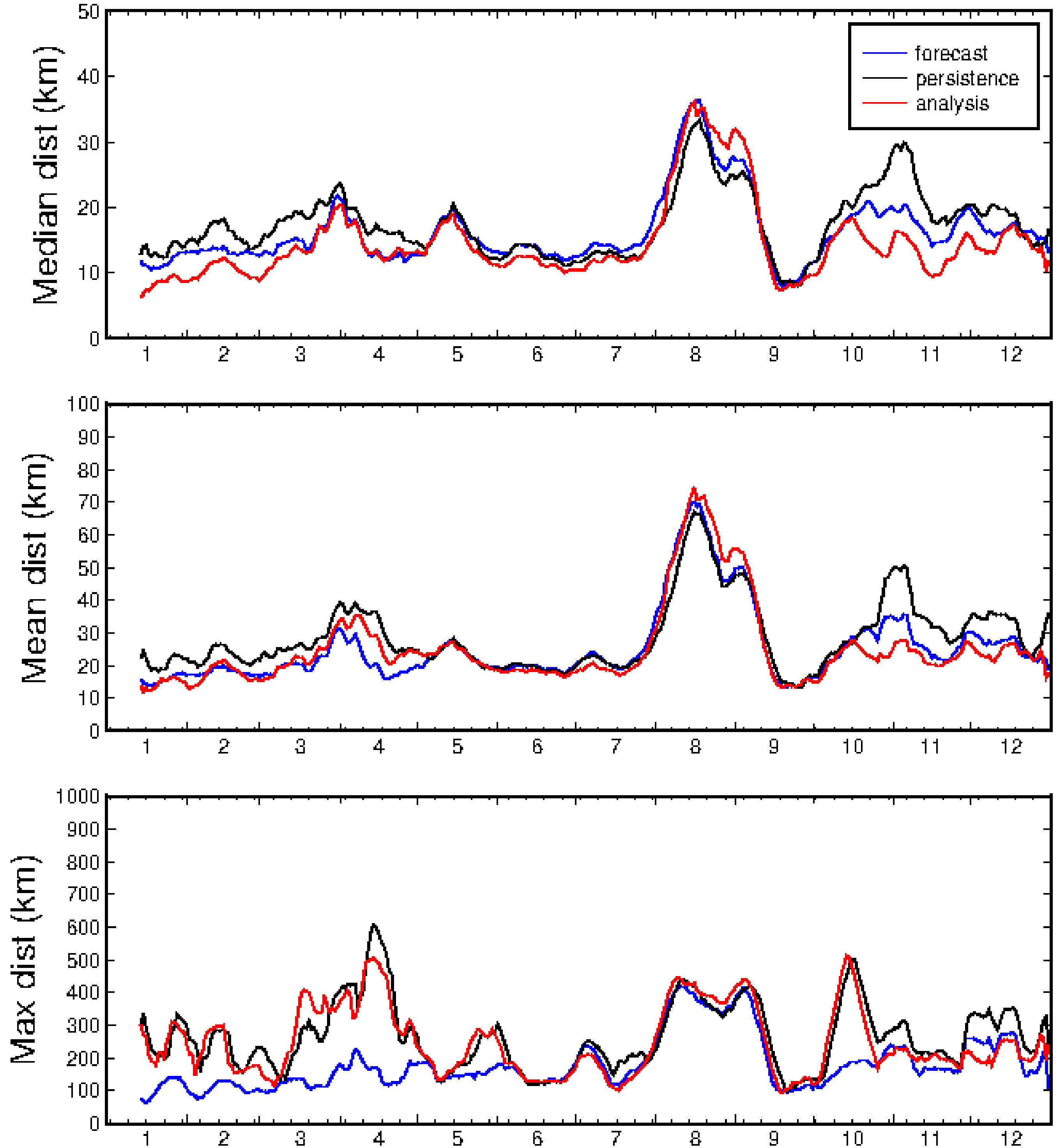
<u>Meaningful verification</u>: intuitive verification statistics, provides a <u>distance in km!</u>

No interpolation of the forecast nor of the obs is required

Distance to Ice Edge

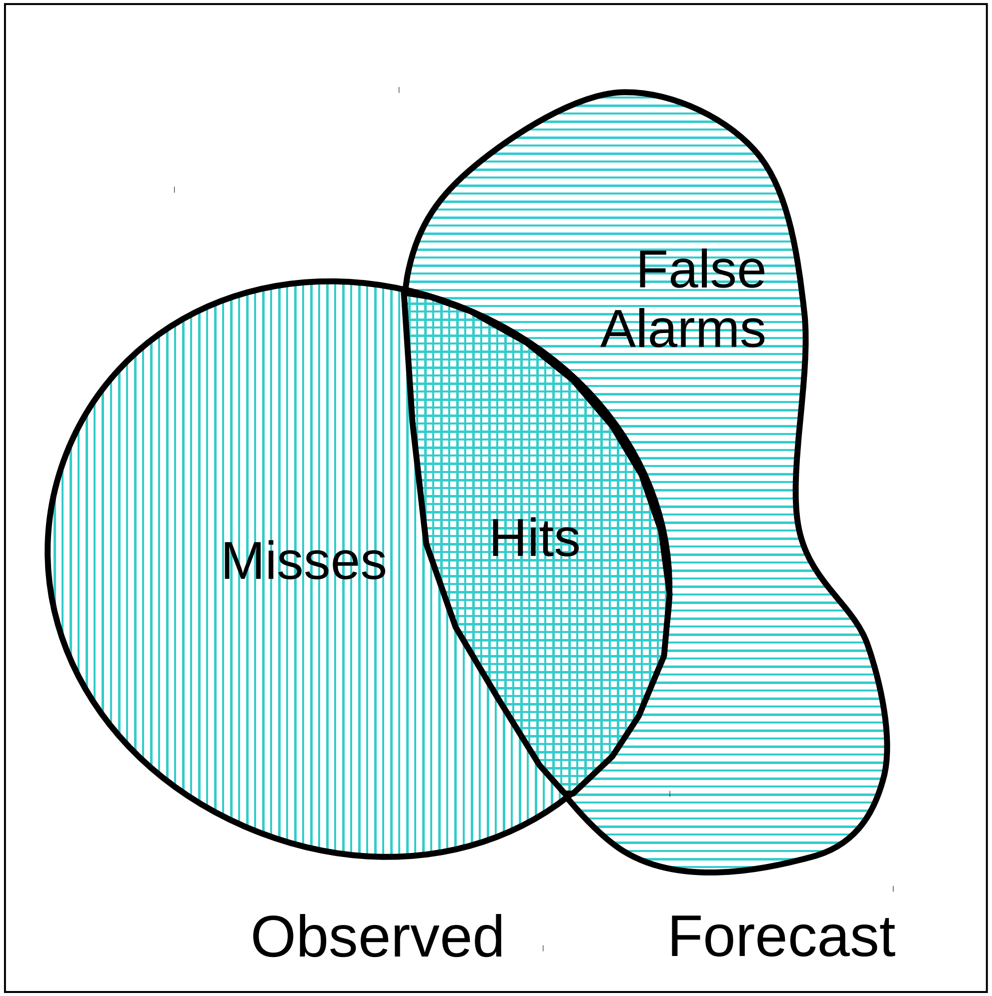
Image and analysis by JF Lemieux (MRD-ECCC)

1. Thresholding: identify forecast and obs ice edges.
2. For each RIPS ice-edge pixel (with dist larger that 50km from coast), evaluate the distance between ice edges.
3. Consider median, mean and max distance (similarly to partial, modified and Hausdorff, but solely forward distances).



Extras 3
RIPS vs IMS, 2011
Categorical Verification

Traditional categorical scores evaluated from contingency tables



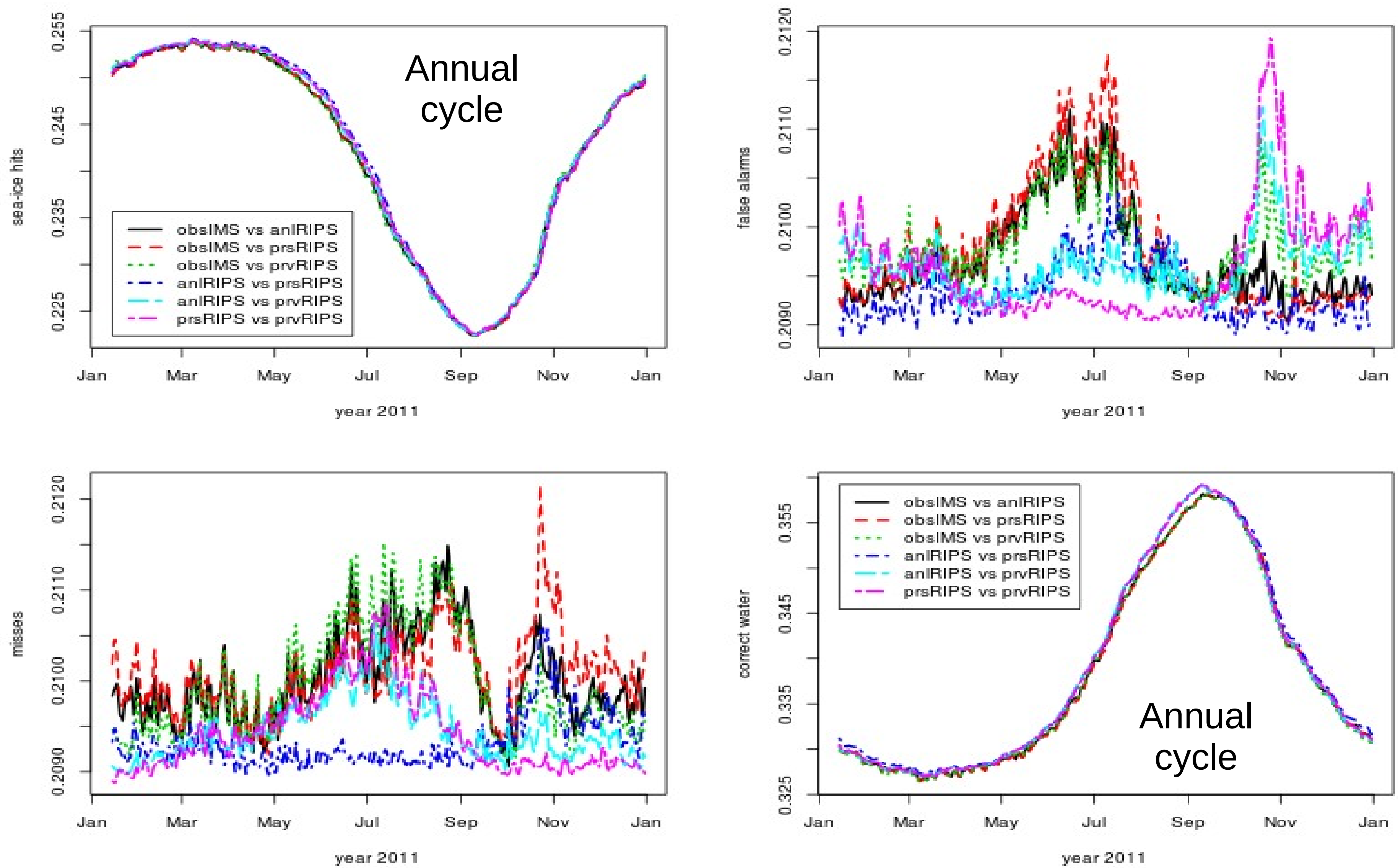
		Observed	
		yes	no
Predicted	yes	*hits*	*false alarms*
	no	*misses*	*nils*

$$FBI = \frac{hits + false\ \ alarms}{hits + misses}$$

$$PC = \frac{hits + nils}{total}$$

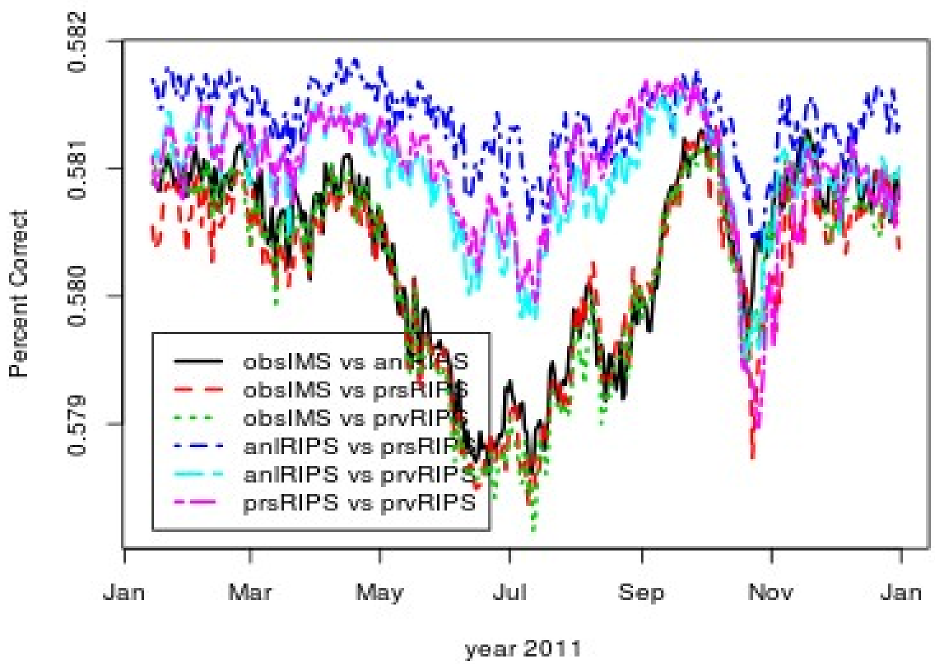
$$HSS = \frac{hits + corr.neg\ - hits_{random} - nils_{random}}{total\ - hits_{random} - nils_{random}}$$

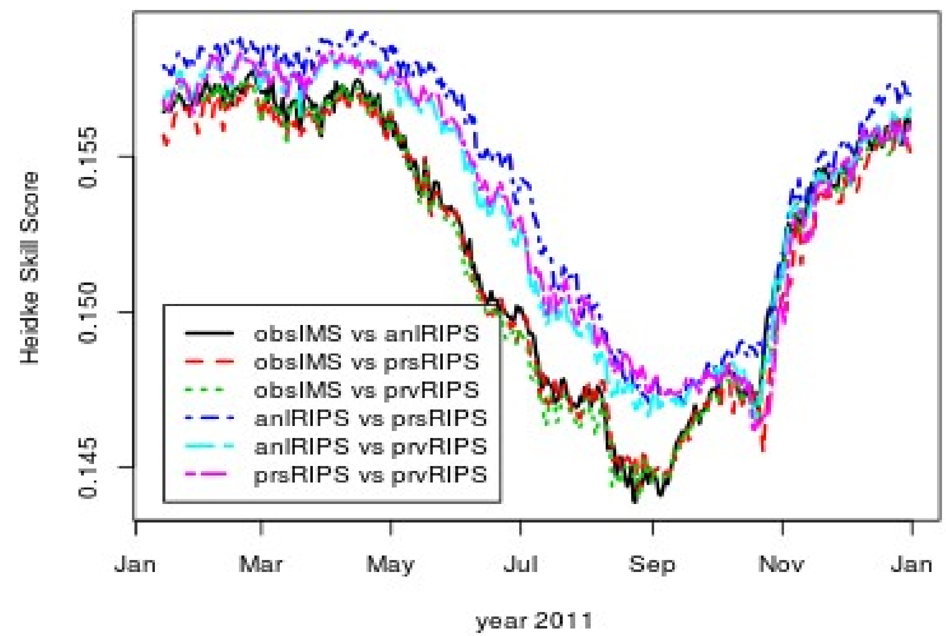
Contingency table entries, RIPS vs IMS

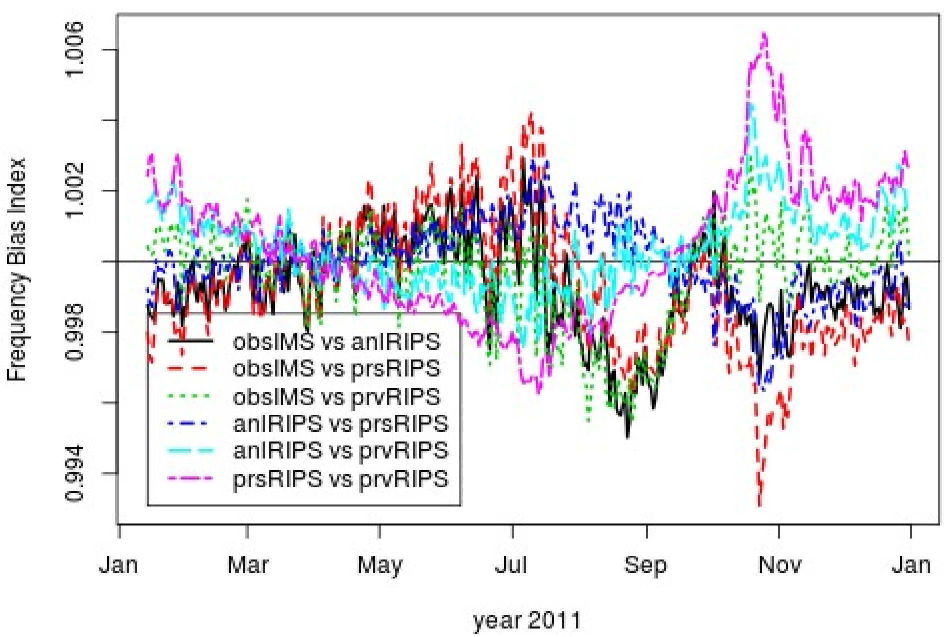


Note the range: sea-ice hits and correct water ~ 0.3;
misses and false alarms ~ 0.03

Categorical scores, RIPS vs IMS

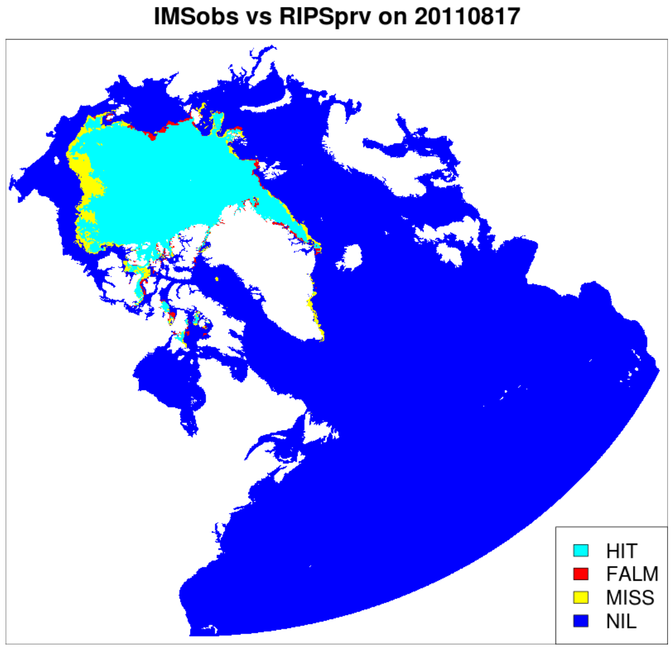


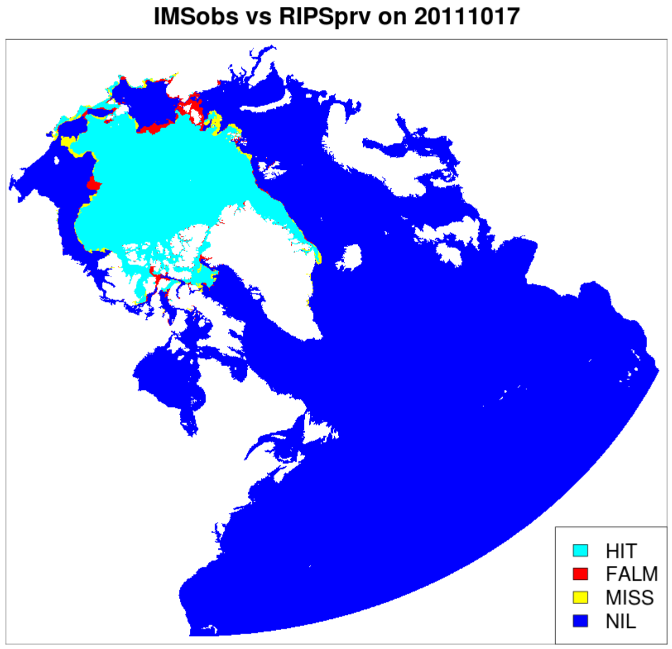




The annual cycles of hits and nils compensate each other. The PC is mostly affected by false alarms and misses (range ~ 0.03).

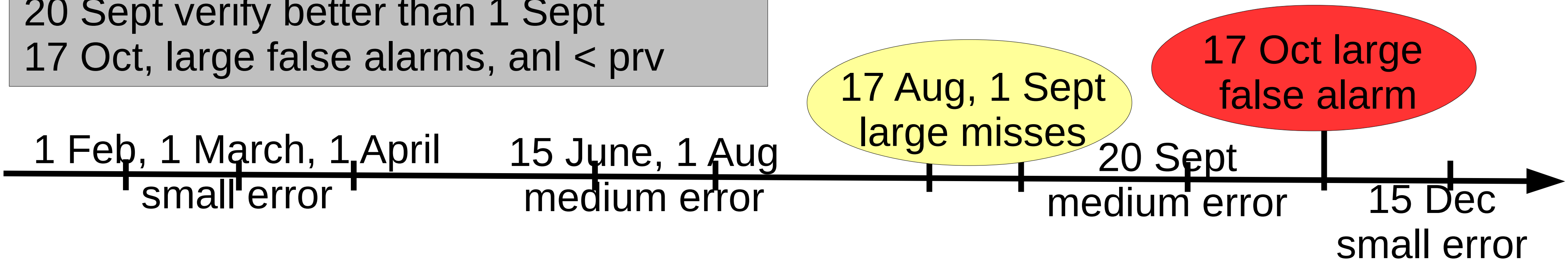
The HSS annual cycle is dominated by the hits, with influences of the false alarms and misses.

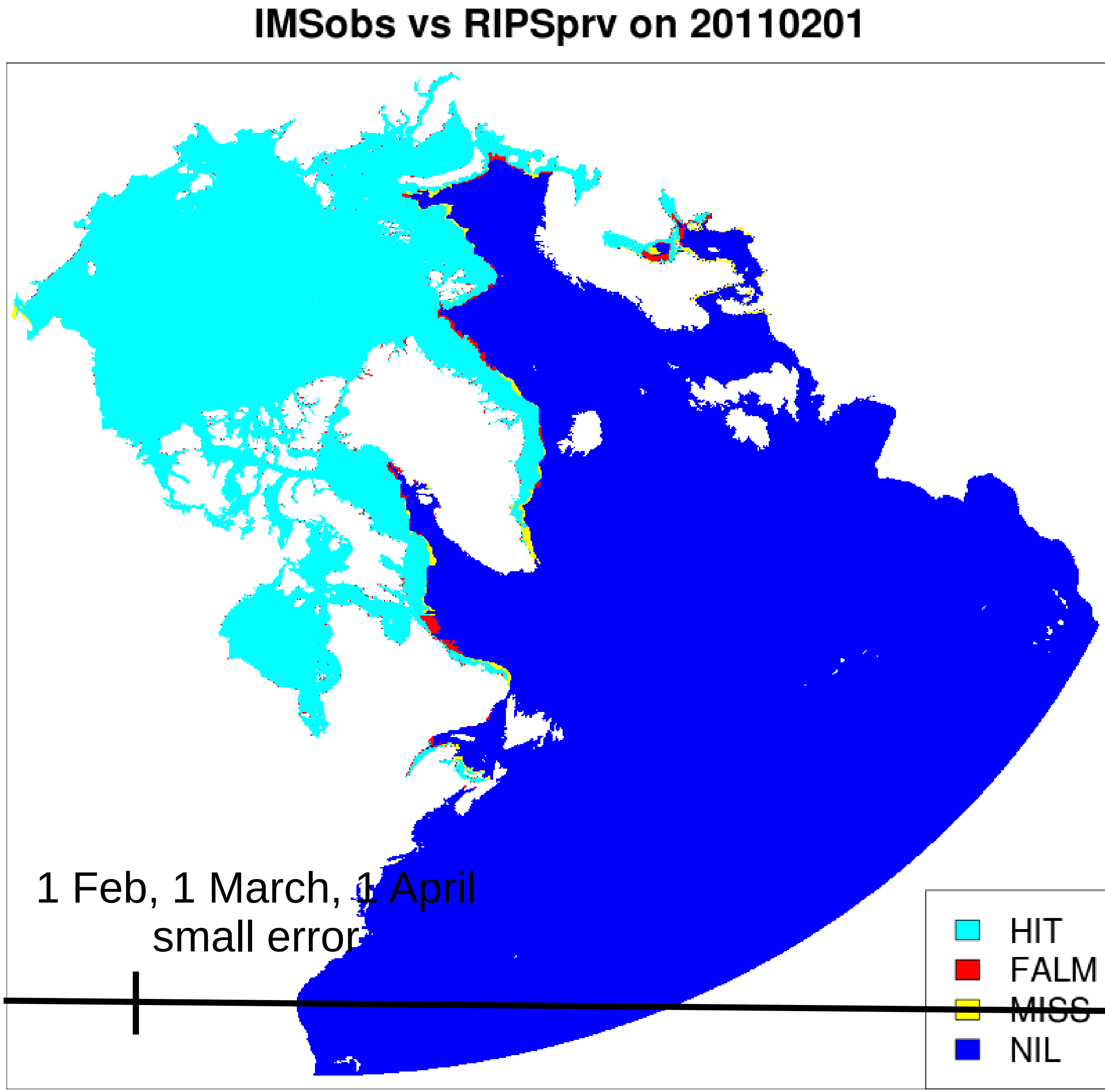




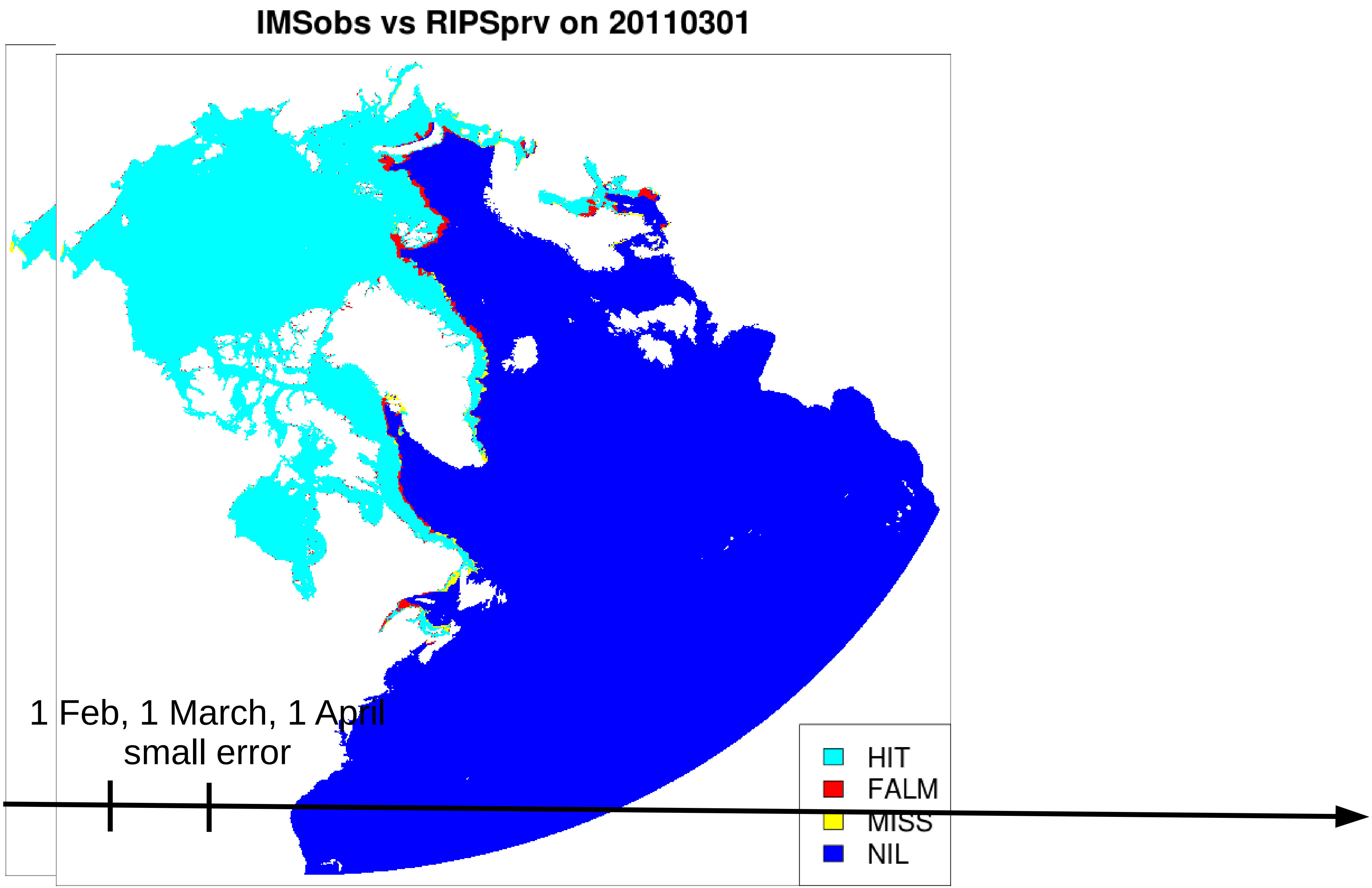
Extras 4
RIPS vs IMS, 2011 annual cycle

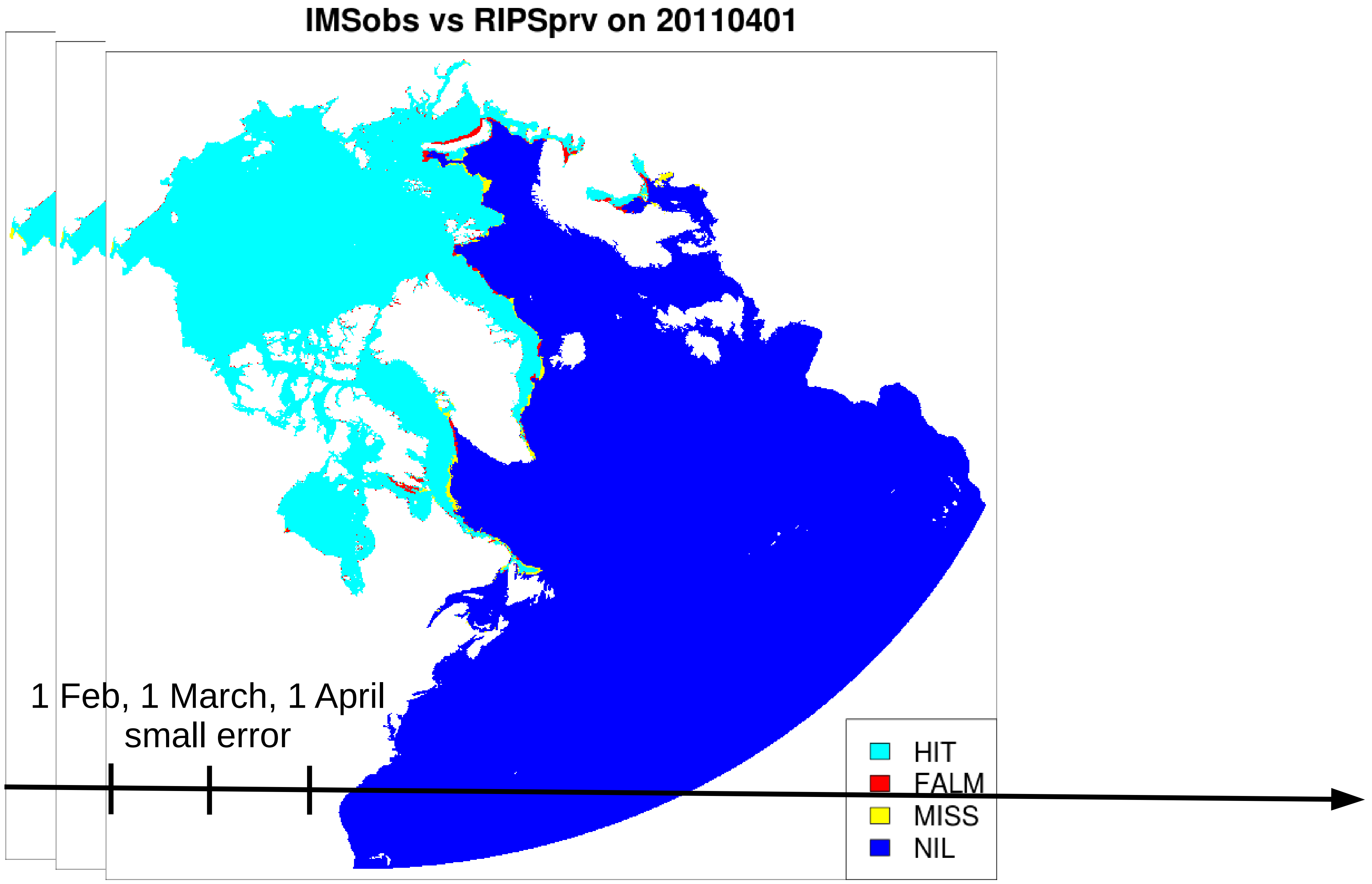
15 Dec – 1 April: small error, anl ~ prv
15 June, 1 Aug: medium error, anl ~ prv
17 Aug, 1 Sept: large misses, anl ~ prv
20 Sept verify better than 1 Sept
17 Oct, large false alarms, anl < prv

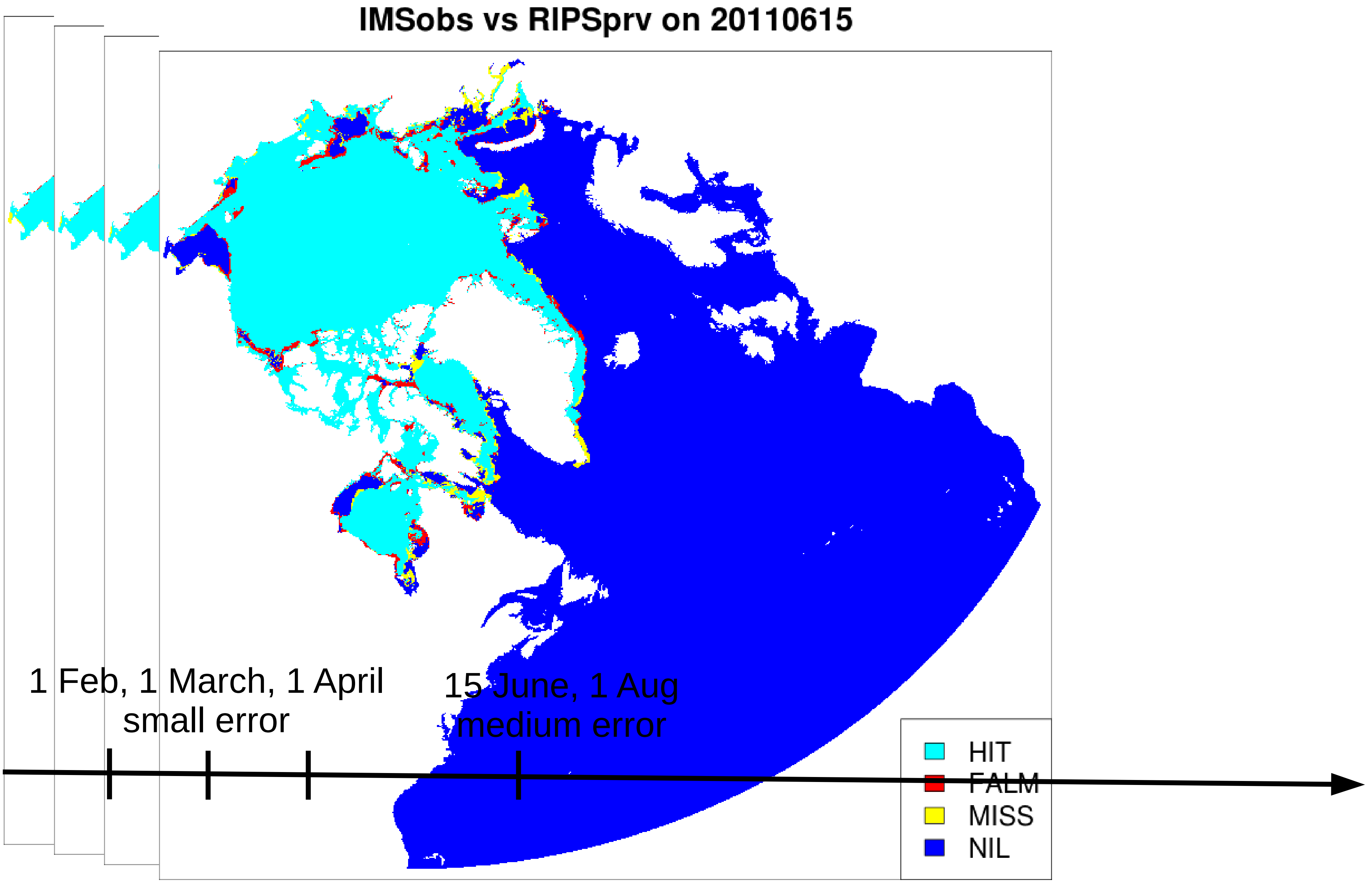


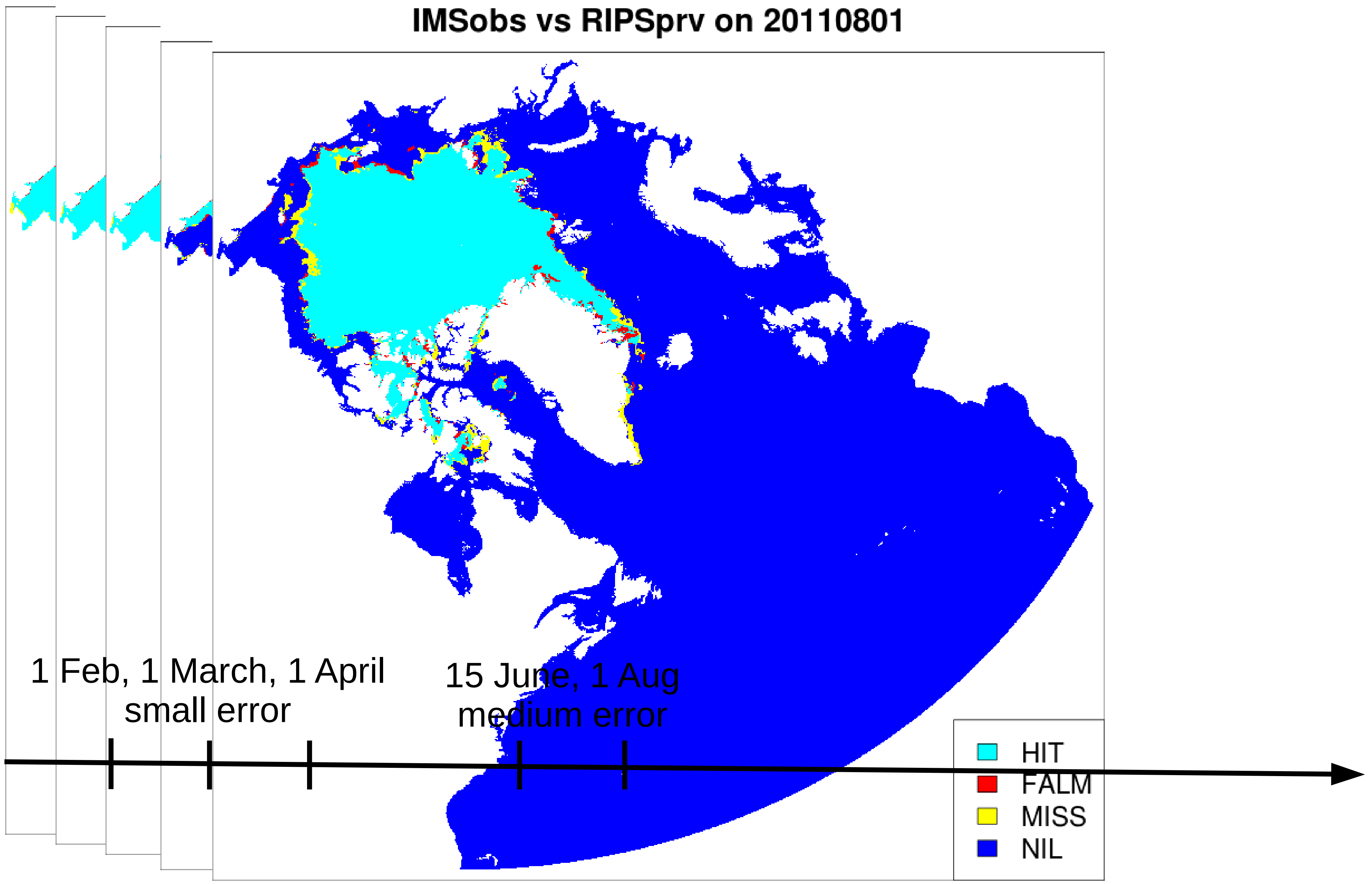


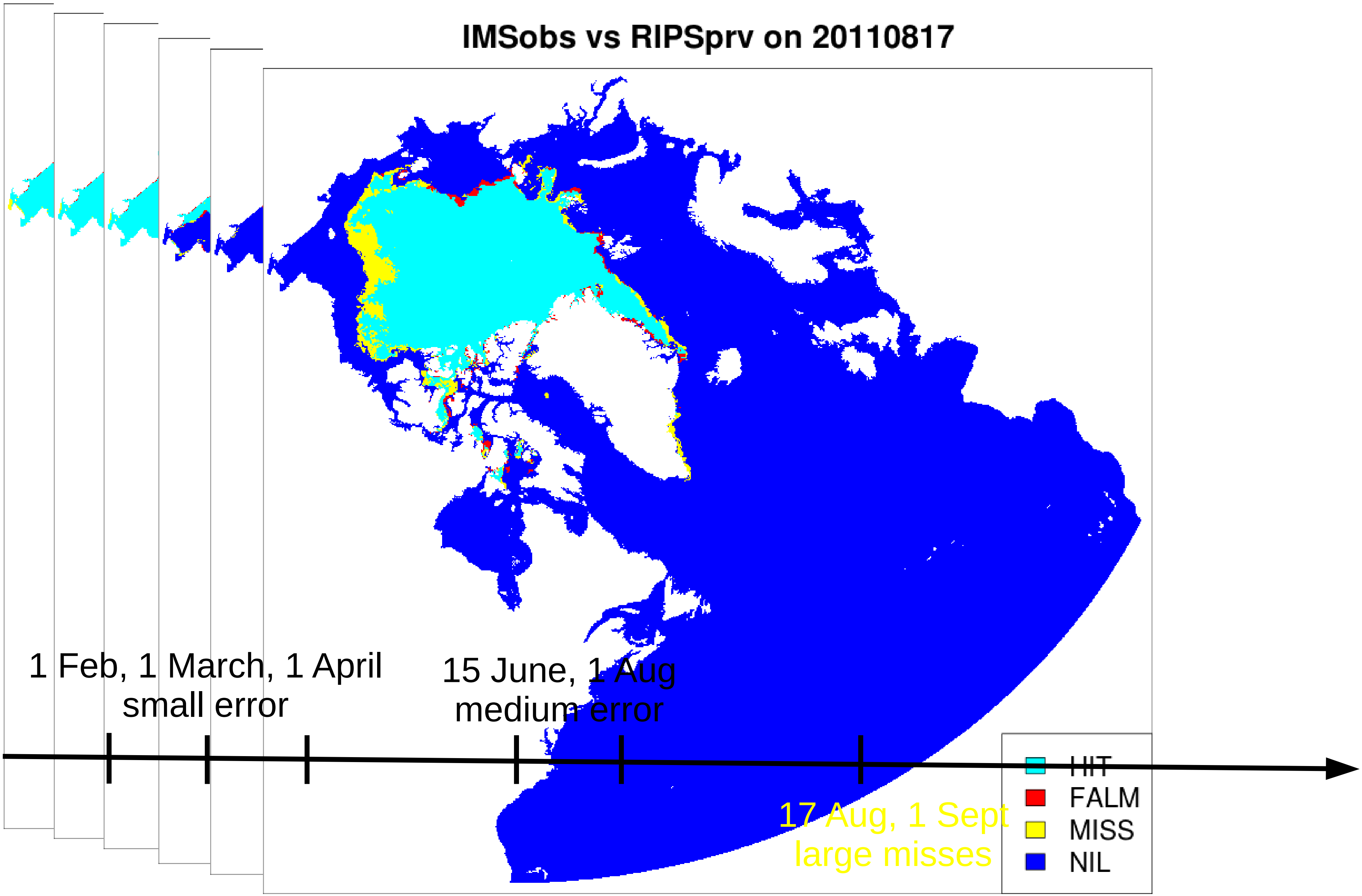
1 Feb, 1 March, 1 April
small error

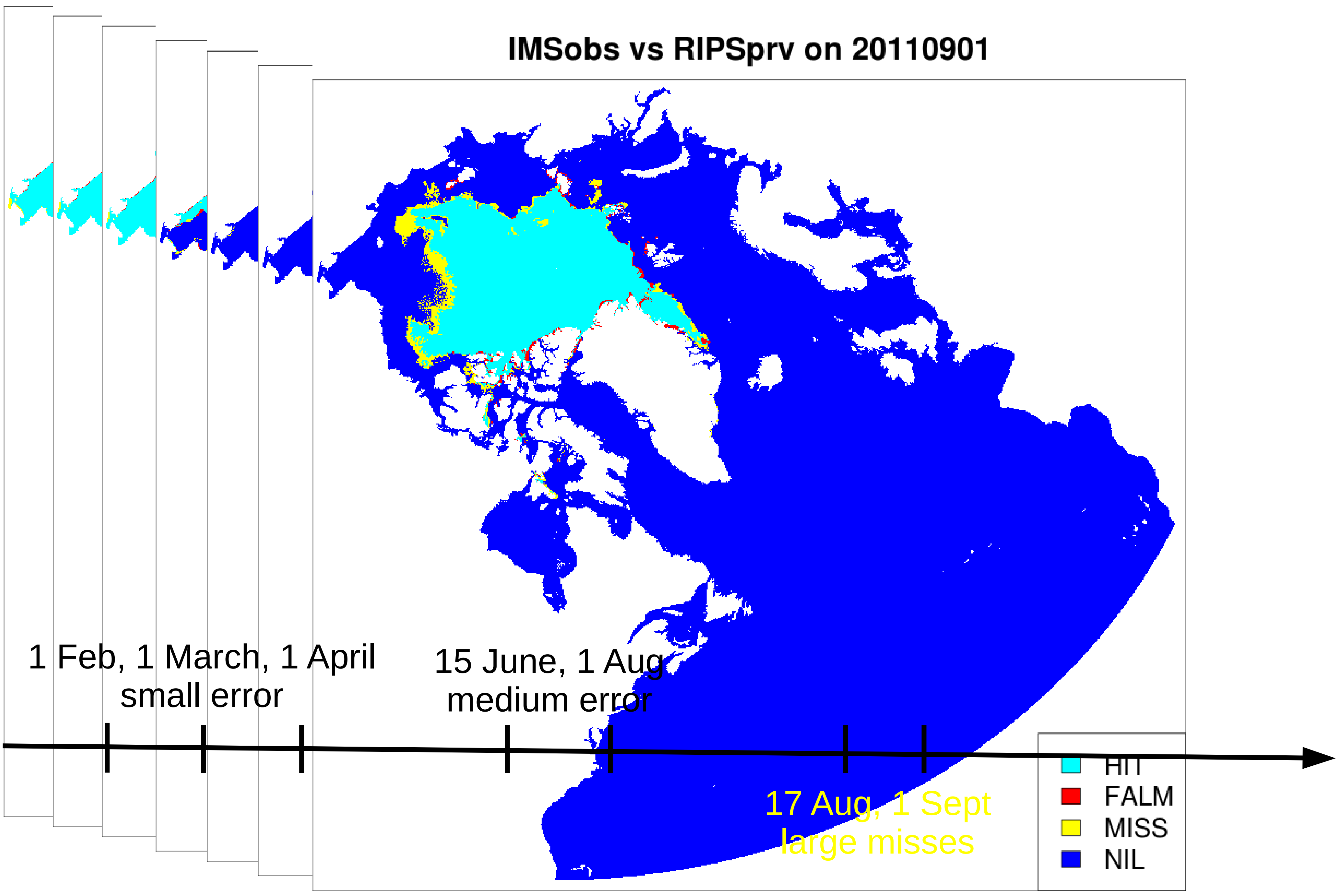

IMSobs vs RIPSprv on 20110301
1 Feb, 1 March, 1 April
small error
HIT
FALM
MISS
NIL

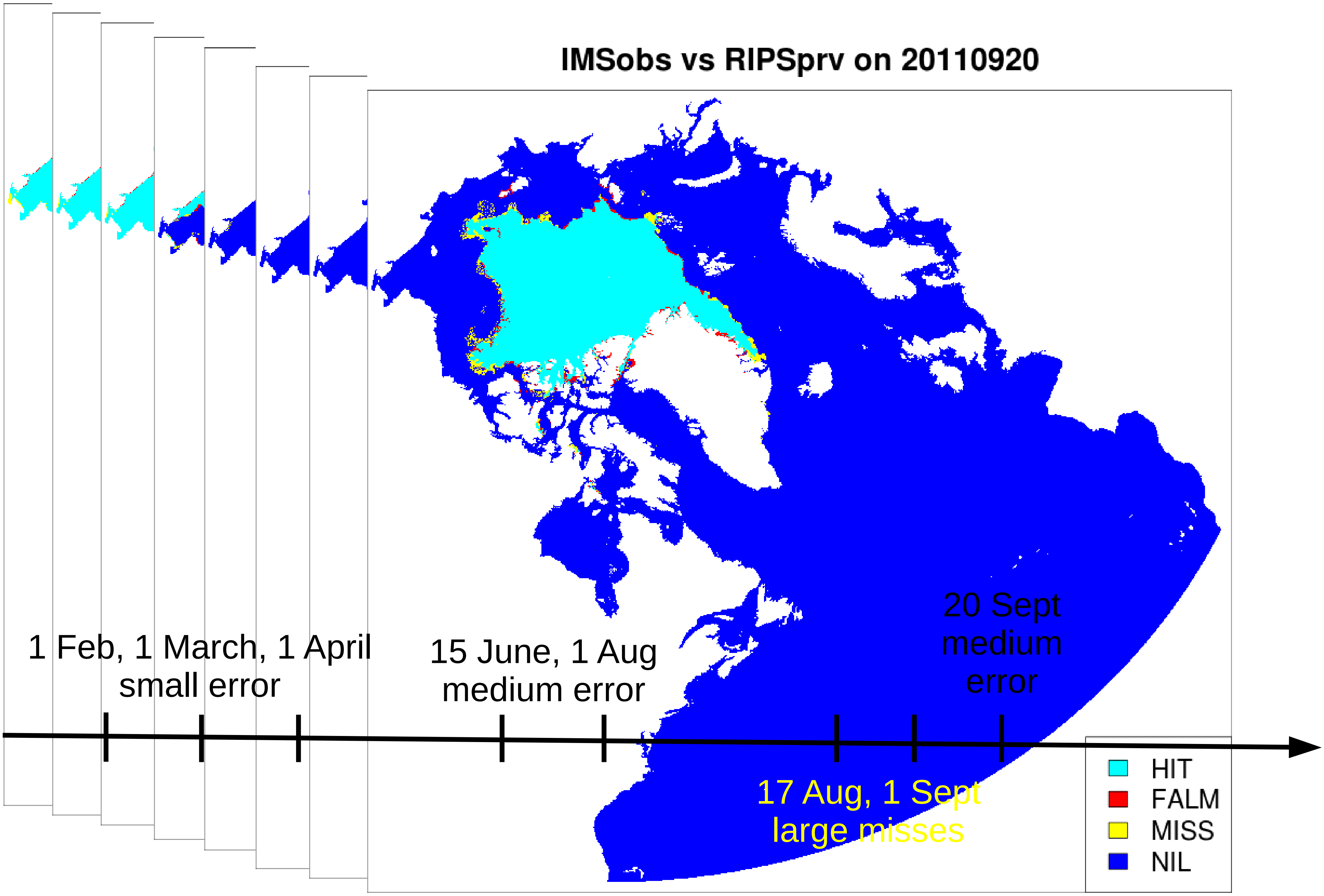

IMSobs vs RIPSprv on 20110401
1 Feb, 1 March, 1 April
small error
HIT
FALM
MISS
NIL

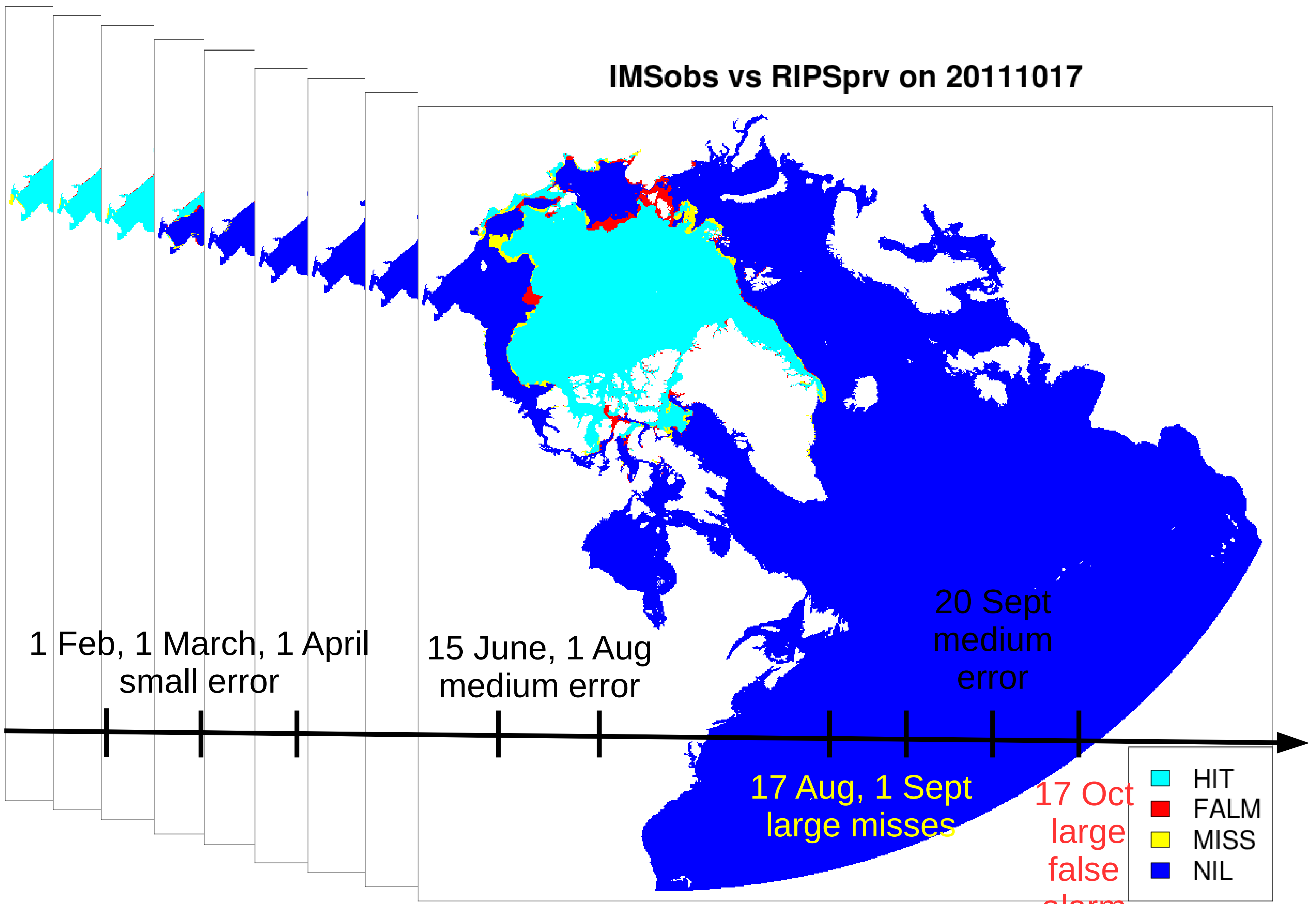

IMSobs vs RIPSprv on 20110615
1 Feb, 1 March, 1 April
small error
15 June, 1 Aug
medium error
HIT
FALM
MISS
NIL

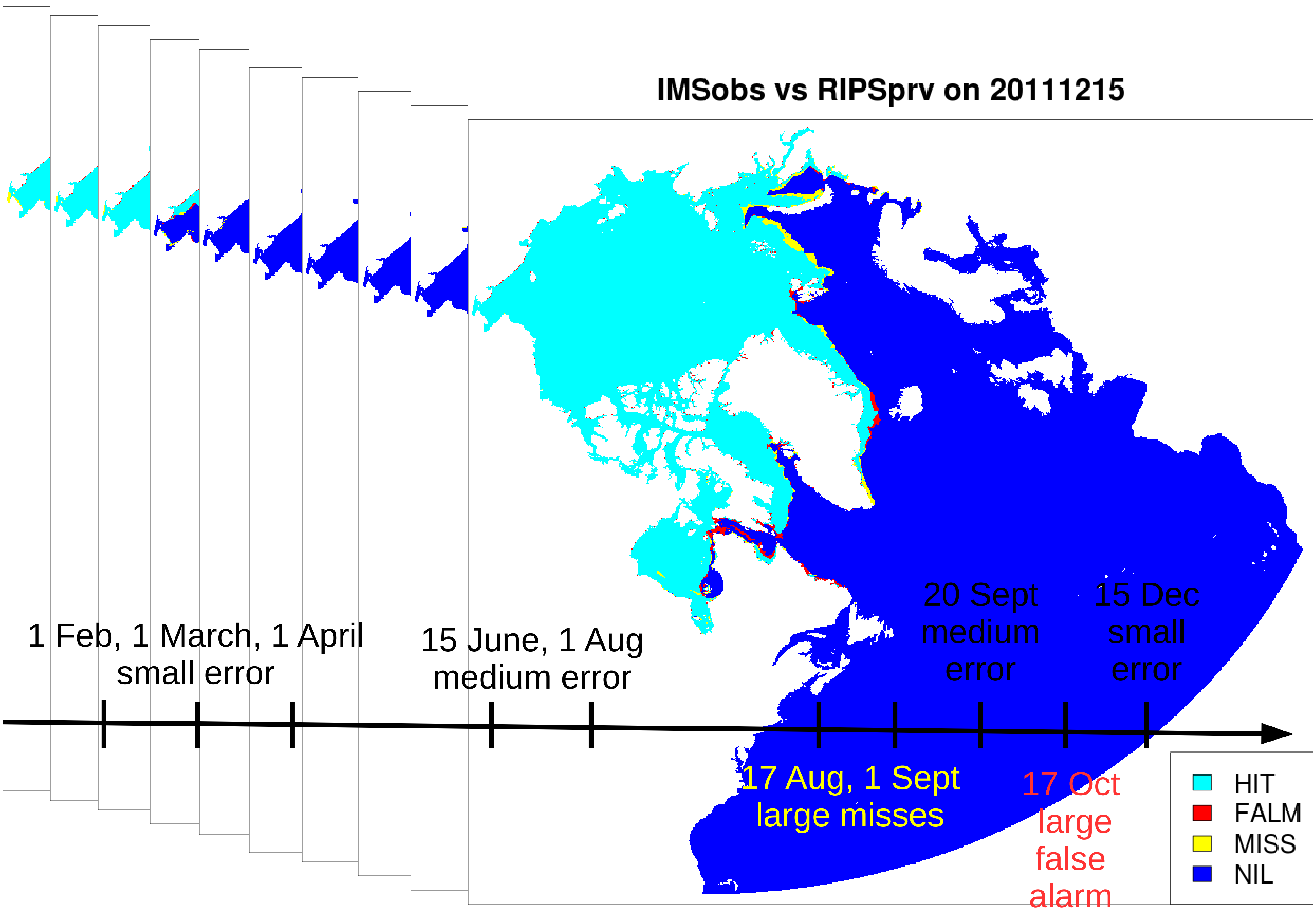

IMSobs vs RIPSprv on 20110801
1 Feb, 1 March, 1 April
small error
15 June, 1 Aug
medium error
HIT
FALM
MISS
NIL


IMSobs vs RIPSprv on 20110817
1 Feb, 1 March, 1 April
small error
15 June, 1 Aug
medium error
17 Aug, 1 Sept
large misses
HIT
FALM
MISS
NIL


IMSobs vs RIPSprv on 20110901
1 Feb, 1 March, 1 April
small error
15 June, 1 Aug
medium error
17 Aug, 1 Sept
large misses
HIT
FALM
MISS
NIL


IMSobs vs RIPSprv on 20110920
1 Feb, 1 March, 1 April
small error
15 June, 1 Aug
medium error
20 Sept
medium
error
17 Aug, 1 Sept
large misses
HIT
FALM
MISS
NIL


IMSobs vs RIPSprv on 20111017
1 Feb, 1 March, 1 April
small error
15 June, 1 Aug
medium error
20 Sept
medium
error
17 Aug, 1 Sept
large misses
17 Oct
large
false
alarm
HIT
FALM
MISS
NIL


IMSobs vs RIPSprv on 20111215
1 Feb, 1 March, 1 April
small error
15 June, 1 Aug
medium error
17 Aug, 1 Sept
large misses
20 Sept
medium
error
17 Oct
large
false
alarm
15 Dec
small
error
HIT
FALM
MISS
NIL