Collection, processing, utilisation and privacy issues of crowdsourced data with a focus on smartphones

Kasper Hintz
kah@dmi.dk
research.dmi.dk

DMI
Danish Meteorological Institute

Innovation Fund Denmark

High Resolution Hirlam Limited Area Model

EWGLAM and SRNWP meetings, Sofia 2019

Photo by John Fowler

20 August 2007
Photo: Casper H Petersen

2 July 2011
Photo: Unknown

4 September 2015
Photo: Jens Larsen

Motivation: better ‘nowcasting’ forecasts and products for our meteorologists by provide a forecast capability and service to warn rapidly developing, extreme weather events.

This means rapid update cycles and better use of current observations and exploits new high-resolution observation types (e.g., crowdsourced data)

Smartphone Pressure Observations

- Motivation
- Why are there barometers in smartphones? (with an example)

Link with Applications

- Collection of Smartphone data
- Data policy and privacy constraints

Assimilation in the HARMONIE system

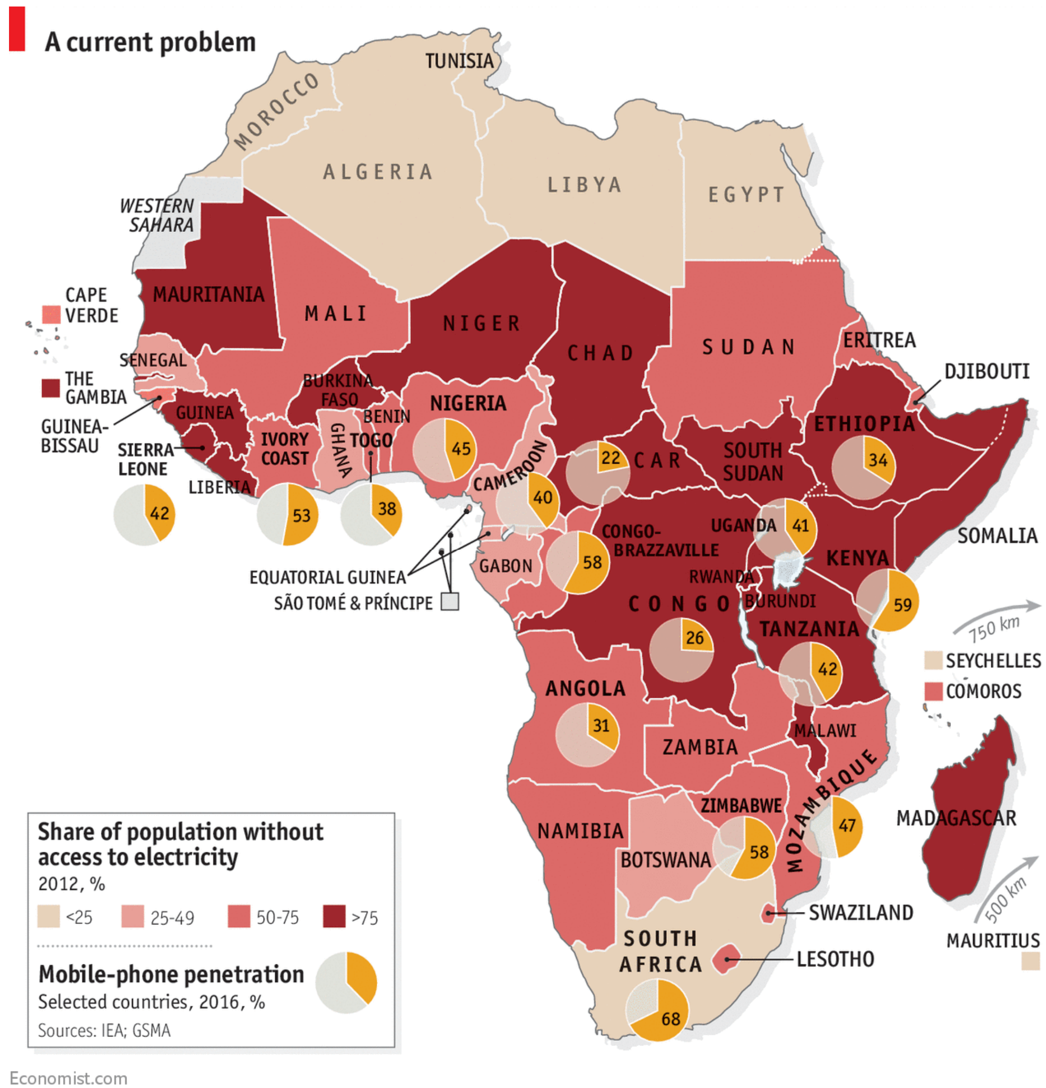
- Results from 3DVar experiment
- Future aspects and experiments (COMEPS)

Another motivational point

According to "*The Economist*" (2016)*:

"In much of sub-Saharan Africa, mobile phones are more common than access to electricity."

"In 2016 two-fifths of people in sub-Saharan Africa had mobile phones. Their rapid spread has beaten all sorts of odds. In most African countries, less than half the population has access to electricity. In a third of those countries, less than a quarter does. Yet in much of the continent people with mobile phones outnumber those with electricity, never mind that many have to walk for miles to get a signal or recharge their phones' batteries."

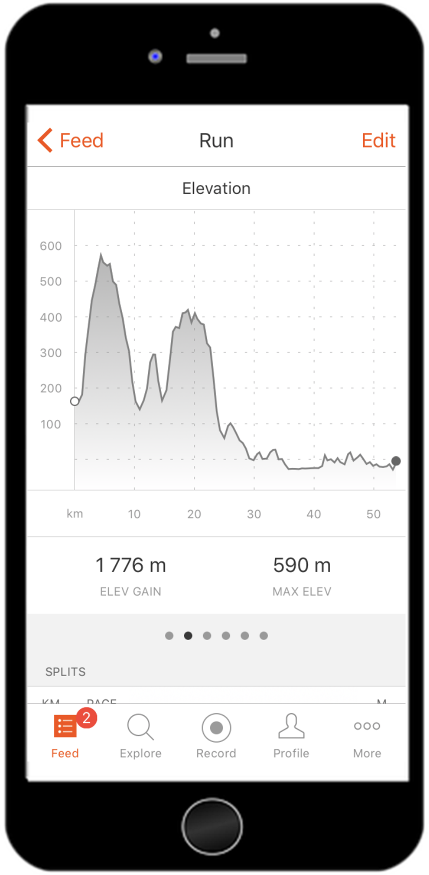


*https://www.economist.com/graphic-detail/2017/11/08/in-much-of-sub-saharan-africa-mobile-phones-are-more-common-than-access-to-electricity

Smartphone Pressure Observations

Most smartphones measures the atmospheric pressure

- Used to monitor changes in altitude and acquire a fix of the location of the device faster.

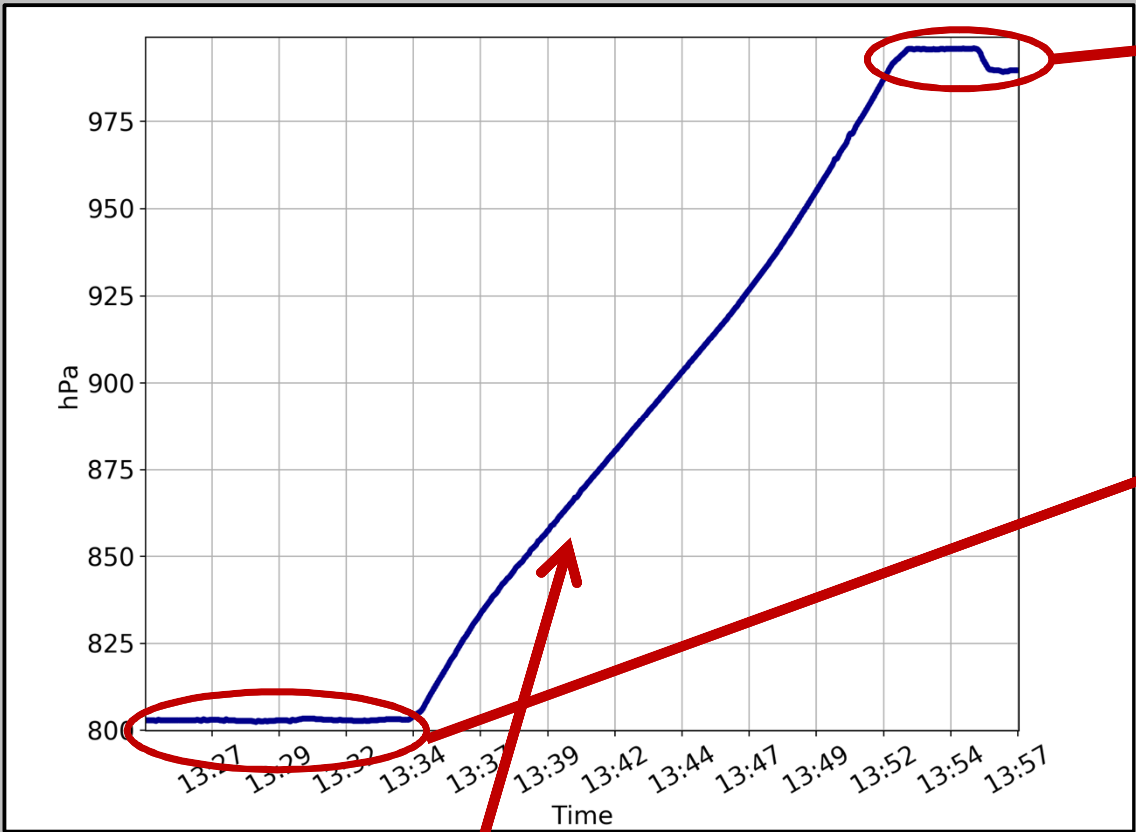


So what?

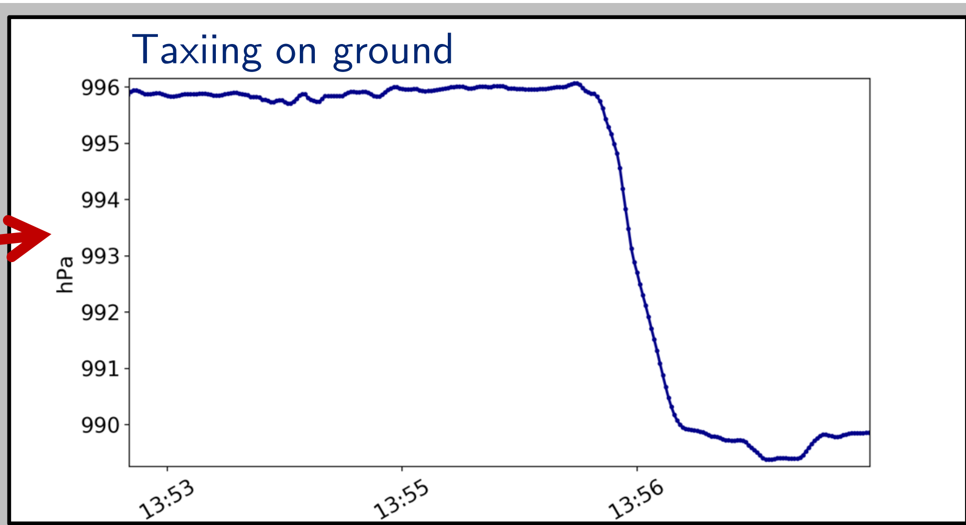
- Pressure is an essential variable in NWP and are being assimilated from conventional sources today.
- Can potentially also be used for verification and/or nowcasting purposes on convective scales.

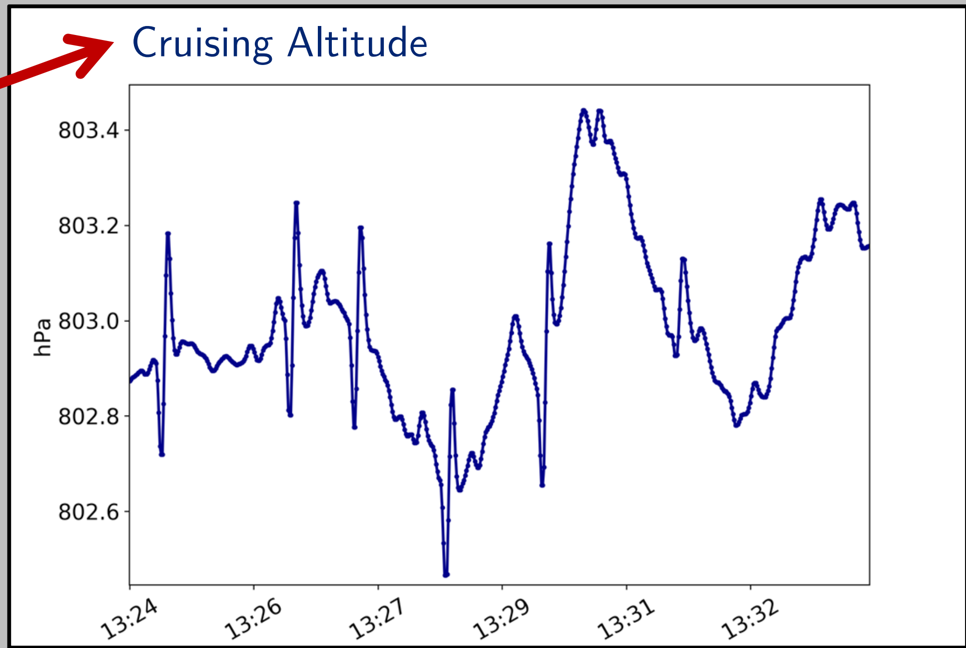
- ***But the GPS altitude is very inaccurate***

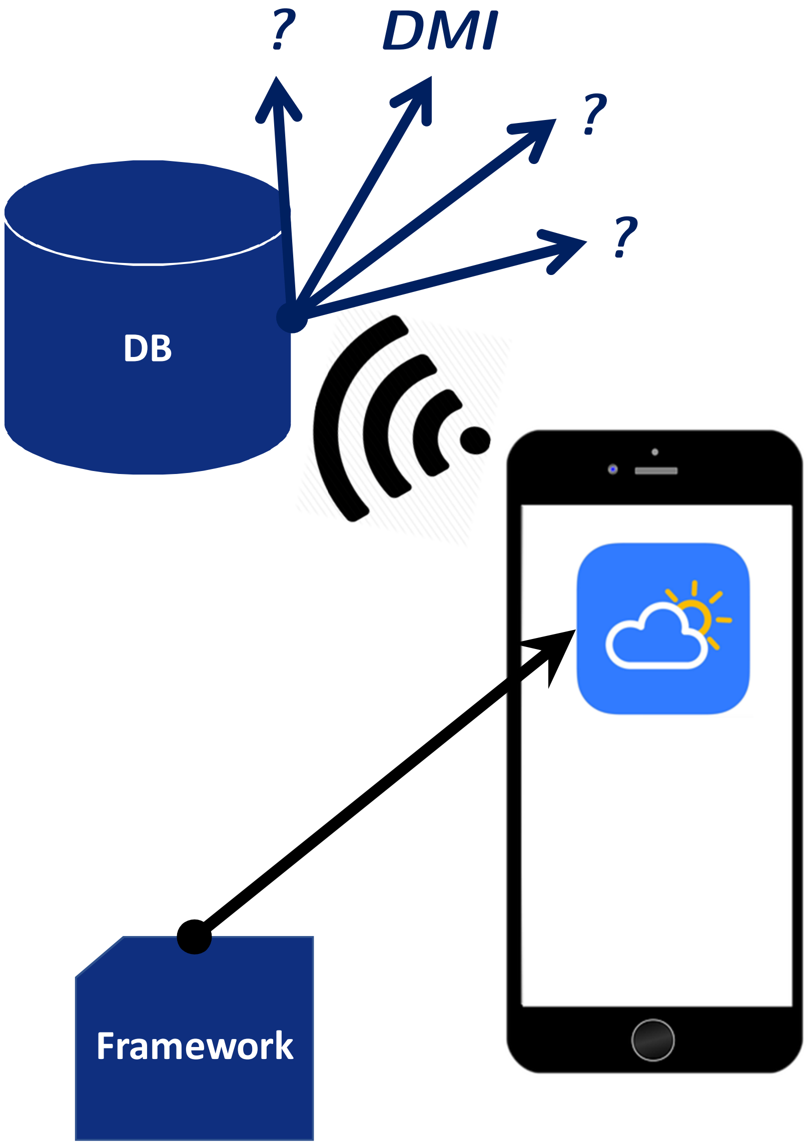
Pressure measured by a smartphone during flight



Descent to Vienna Airport





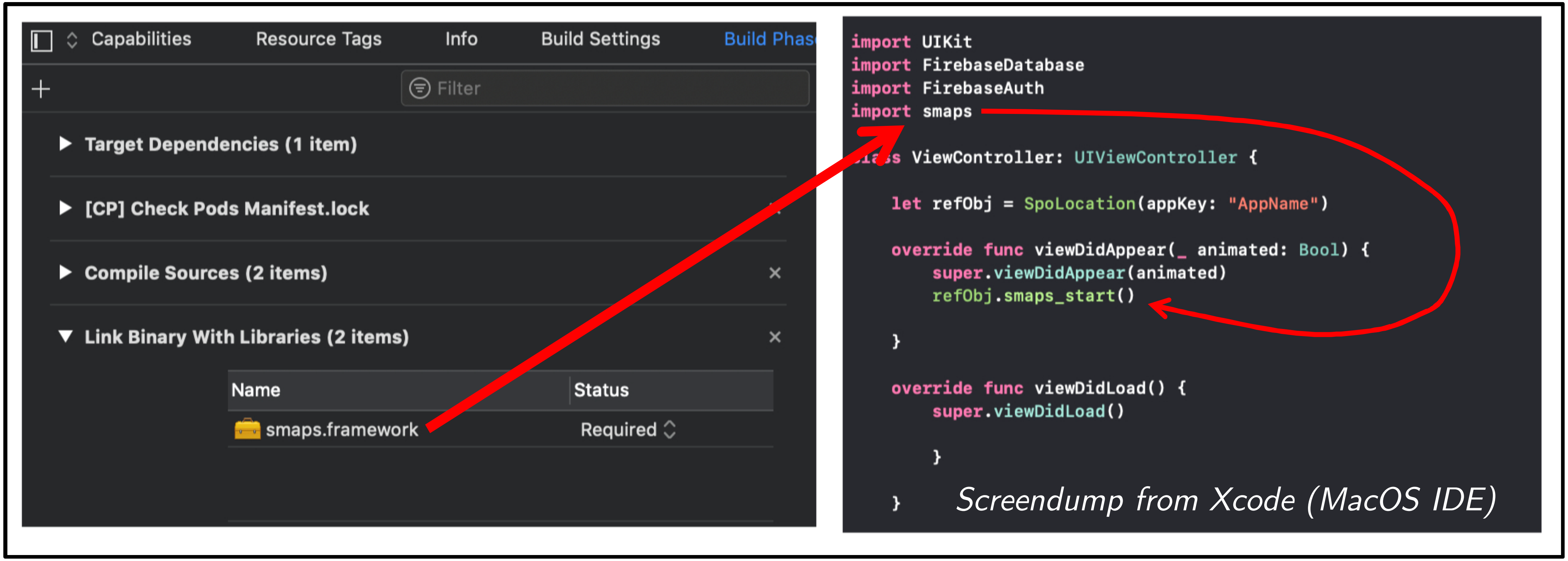


We do *not* develop and maintain a full-stack app with a main focus on collecting observations.

We ***do*** develop and maintain a framework with the sole purpose of collecting observations.

Objective: Keep data processing in the meteorological community and collaborate on data collection between meteorological services.

Basic example of starting observation collection



Screendump from Xcode (MacOS IDE)

There is still room for improvement and we welcome all collaboration on this.

kah@dmi.dk | hev@dmi.dk

It is still unknown if extra (meta)data from the smartphones can be used for quality control and/or correction;

McNicholas and Mass (2018) showed excellent results indicating, that this is the case.

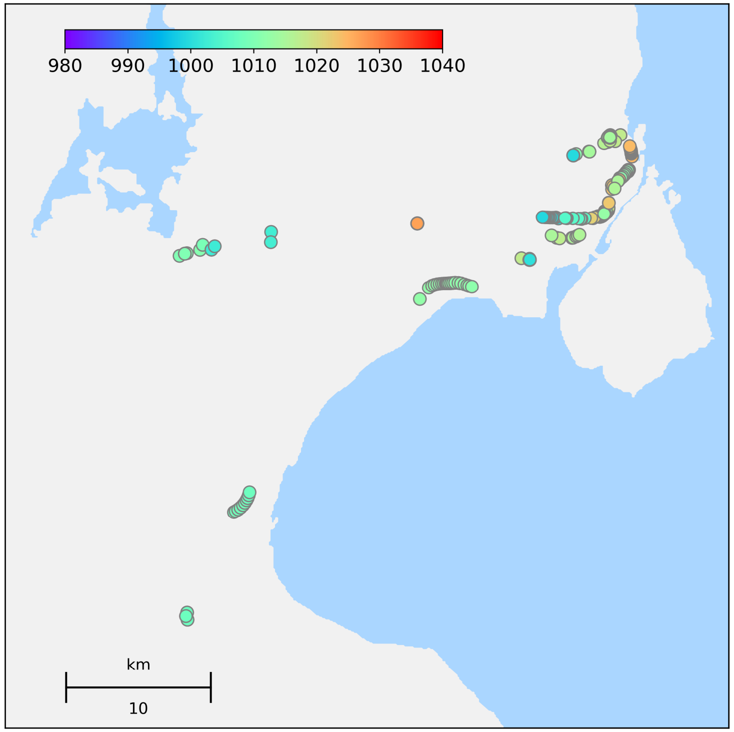
Therefore the following is collected if available:

- **Pressure**
- **Latitude**
- **Longitude**
- **Altitude**
- **Timestamp**
- **User ID**
- **Acceleration** in three dimensions
- **Speed** of device
- Horizontal **Accuracy**
- Vertical **Accuracy**
- **σ** calculated on the phone directly

Furthermore, the following is appended:

- **Residual** (observation-background)
- **DHM** (Danish Terrain Model) height at location

Makes data personal and hence processing must comply with the GDPR act from EU.



Observations from a single unique device over 4 weeks.

Some comments on privacy issues

Currently users are asked for permission to fetch data while the app is open,
(we are working on improving this)

Asking for consent can introduce problems later:

- Permissions must be collected, registered and can be cancelled.
- If the intended data usage changes you have to ask for permission again.
- Can give bias in data (*does properly not apply for NWP*)

Fully anonymized data are not governed by GDPR. (A way around GDPR?)

- Requires data minimization and generalization that often 'destroys' the value of the data.

Article 6: ***Consent***, Contract, Compliance with a legal obligation, Vital Interests, ***Public Interest***, Legitimate Interests.

- Consent is only one among multiple reasons for "Lawfulness of processing".

Table 2.1: Overview of collected smartphone observations from 4th of April 2018 to 4th of April 2019. The decline in April 2019 is because only four days are included from this month.

	Total Observations	**Unique Devices**
April (**2018**)	4,256,983	35,974
May	3,155,072	37,672
June	3,131,730	34,142
July	6,857,070	56,450
August	9,928,301	66,143
September	6,179,692	57,692
October	4,534,645	54,607
November	3,146,321	47,072
December	3,748,830	46,219
January (**2019**)	5,343,834	54,570
February	4,027,405	52,628
March	6,814,125	54,336
April	604,664	22,904
Total	**61,728,672**	**149,782**

From K. S. Hintz, (2019)

During the first year ***61,728,672*** observations was collected from ***149,782*** unique devices.

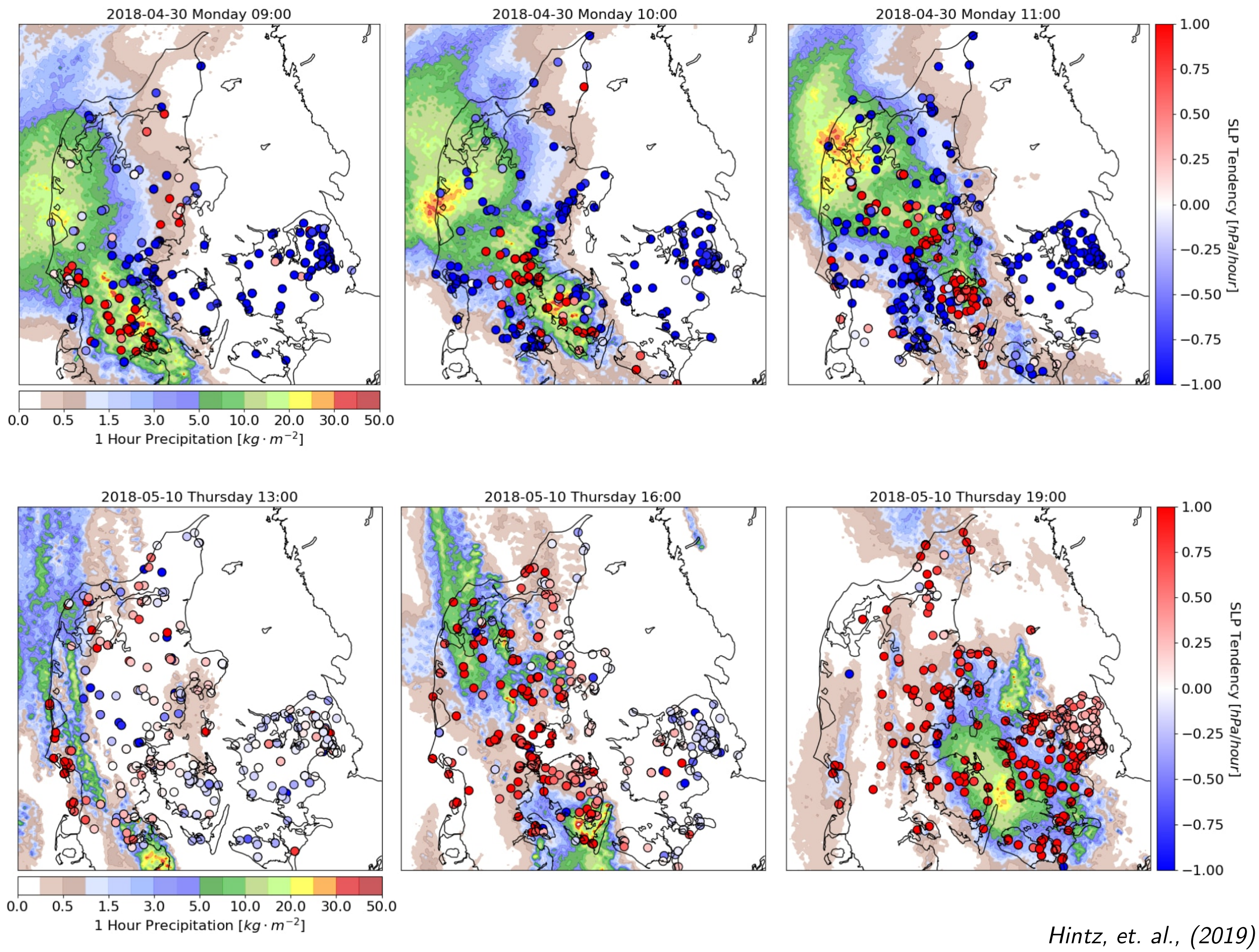
Assuming this is scalable to the rest of Europe, gives about 15 million devices.

10^{-6} € per observation in maintaining costs of database.

Pressure tendencies (raw data) plotted with radar-estimated precipitation.

Frontal zones can be identified directly.

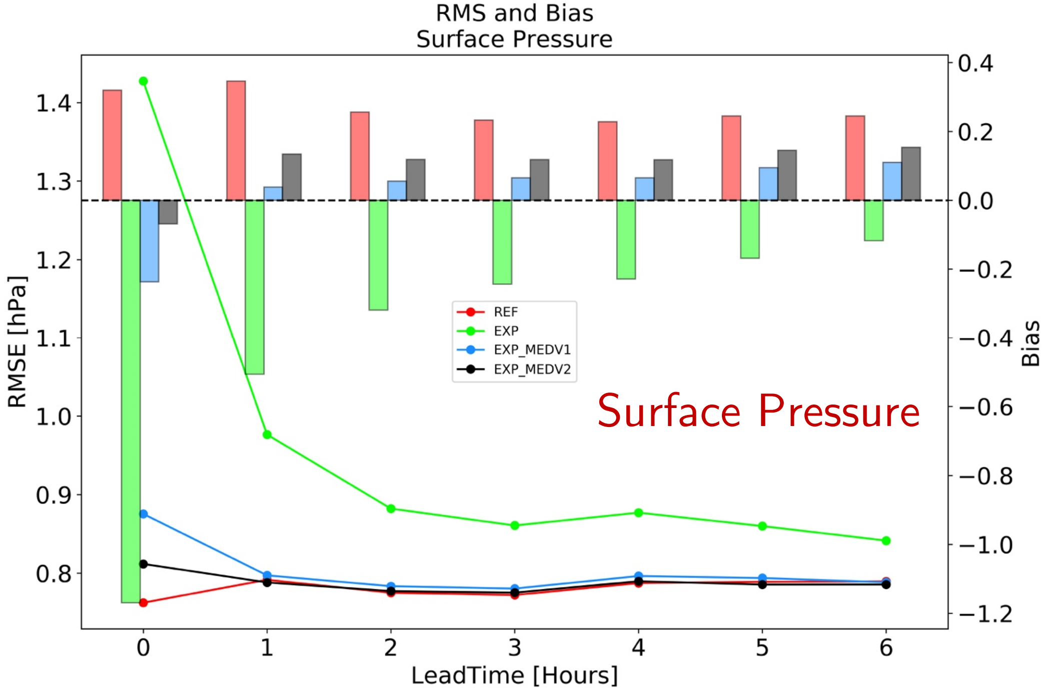
Necessary to ensure only good observations are assimilated without throwing too much away.

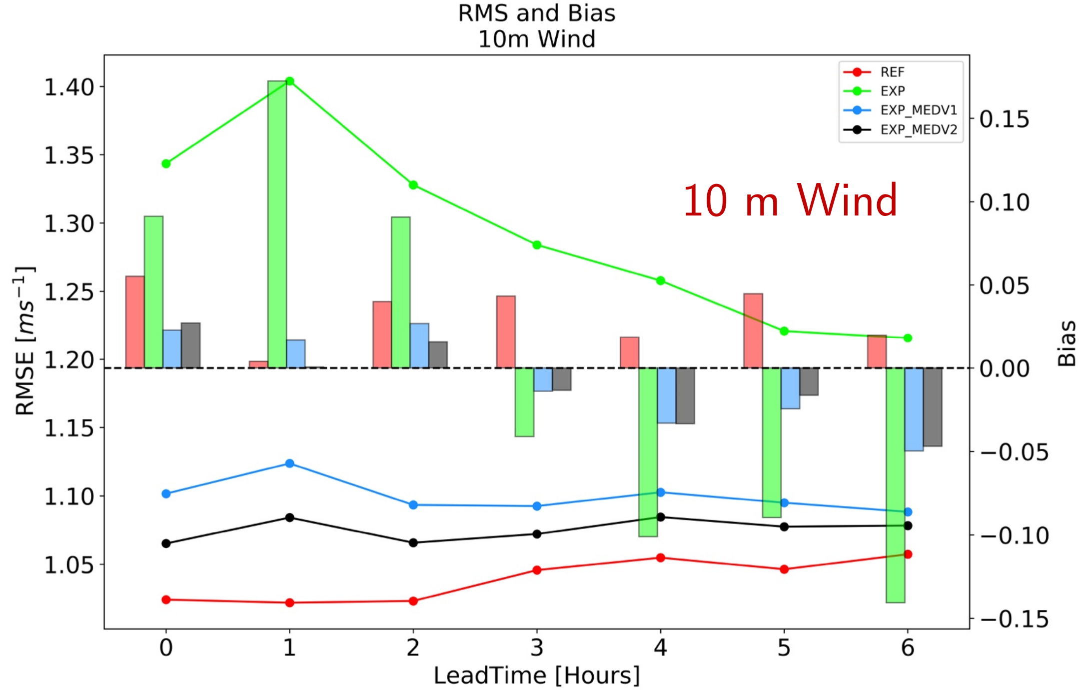


Hintz, et. al., (2019)

NWP Experiments with 3D-Var (Harmonie c40h1)

Date range: 5th May 00 UTC – 10th May 12 UTC. **DA Cycle:** 3 Hours.





Ref: No pressure observations from Denmark

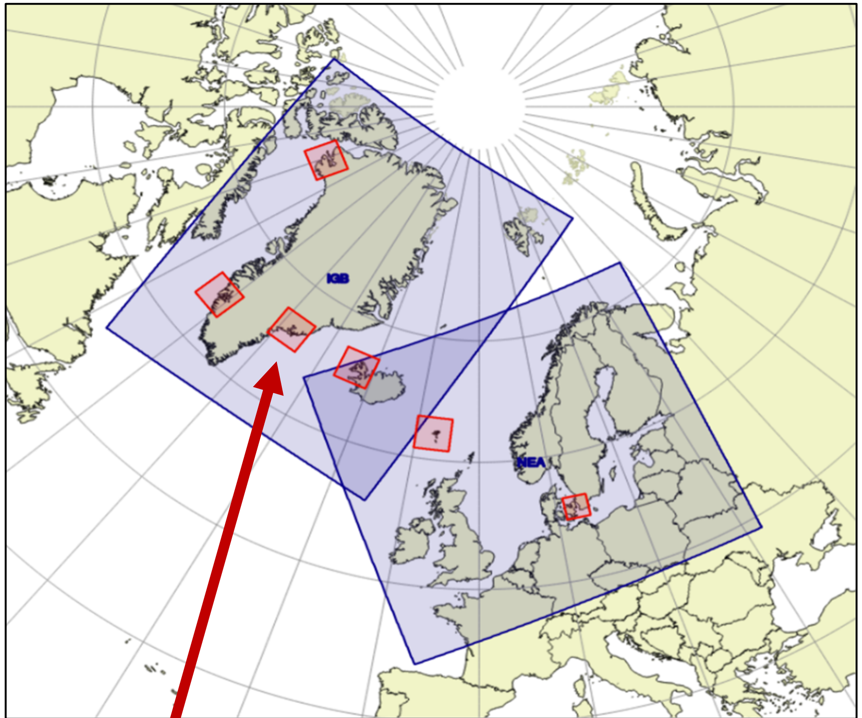
EXP: No filtering of SPO

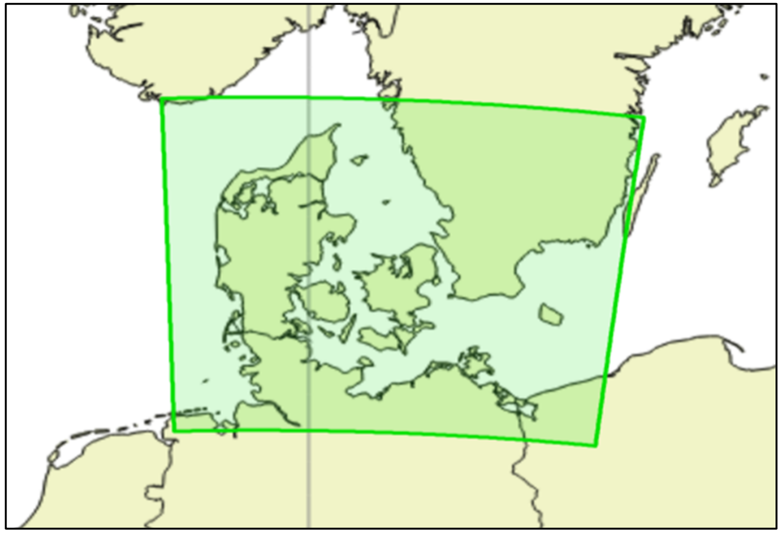
EXP_MEDV1: Filtering with 'loose' median check

EXP_MEDV2: Filtering with 'strict' median check

In another experiment, bias decreased from 0.35 hPa to -0.15 hPa over two months using SPOs.
Screening methods are described in Hintz, et. al, (2019).

"HARMONIE-lite"

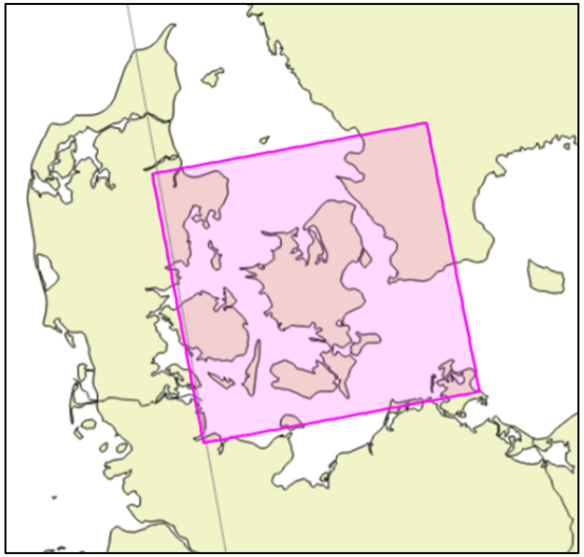




DK750 Model domain
Future Nowcasting
Model.

DA in same resolution
as the model.

DK500 Model domain
(Test setup)



NEA, IGA and COMEPS*
(blue shaded)

(*Continuous Mesoscale
Ensemble Prediction System)

Visit Bent Sass at poster session
for verification results on sub-km
resolution modelling.

Data Assimilation for HARMONIE NWP nowcasting system

Dr Xiaohua Yang, DMI, points out that:

Data assimilation for very high resolution is a necessary capability for NWS to forecast local, small scale weather for extremes.

For some of the high impact weather, the time between first observed phenomena and finish of it are within a few hours.

For NWP end-users, <u>*timeliness*</u> and <u>*consistency with observations*</u> are part of quality indicators.

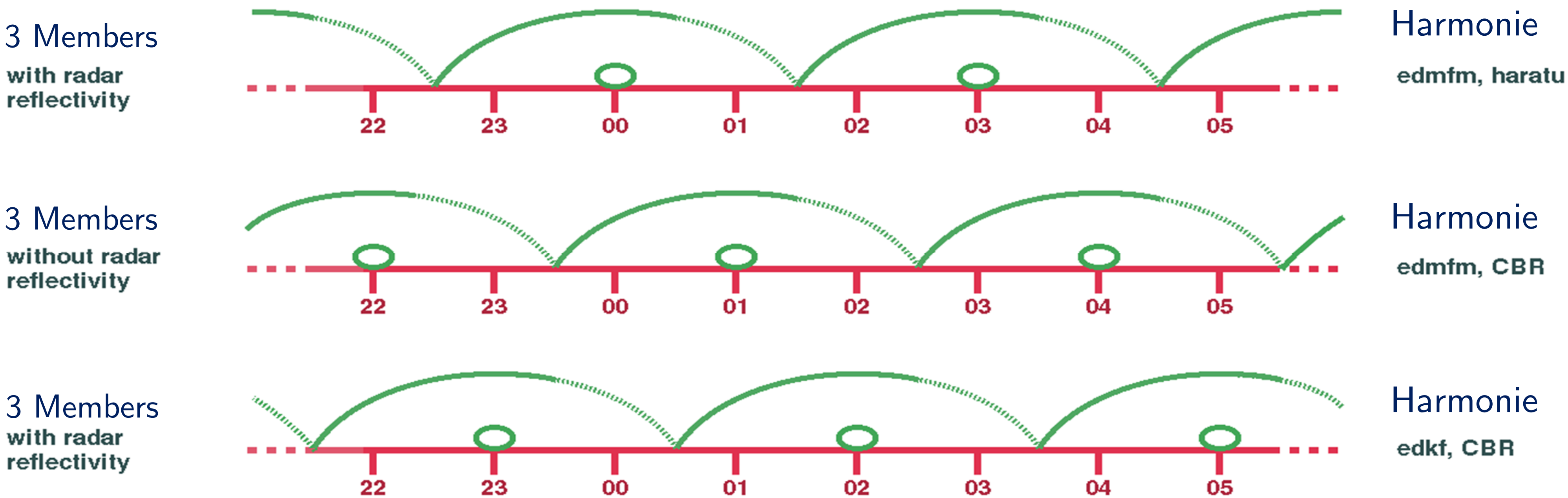
Harmonie-nowcasting shall go to ***sub-km grid scale.***

Observation data used for short range forecasts at DMI

	Production frequency	***Delay time between last observations and forecast at DMI***
Radar advection	Every 10 min	25 min
Nowcasting, Harmonie	Every 1 h	30 min
COMEPS	Every 1 h	2 h 15 min
COMEPS, Nowcasting (Targeted)	Every 10 min	35 min
NEA (Operational Harmonie)	Every 3 h	2 h 30 min – 4 h 30 min
IFS (ECMWF)	Every 6 h	3 h 45 min – 8 h 45 min

Slide credits: ***Xiaohua Yang***, ***DMI***

DMI-COMEPS: Frequent Analysis with Overlapping Windows

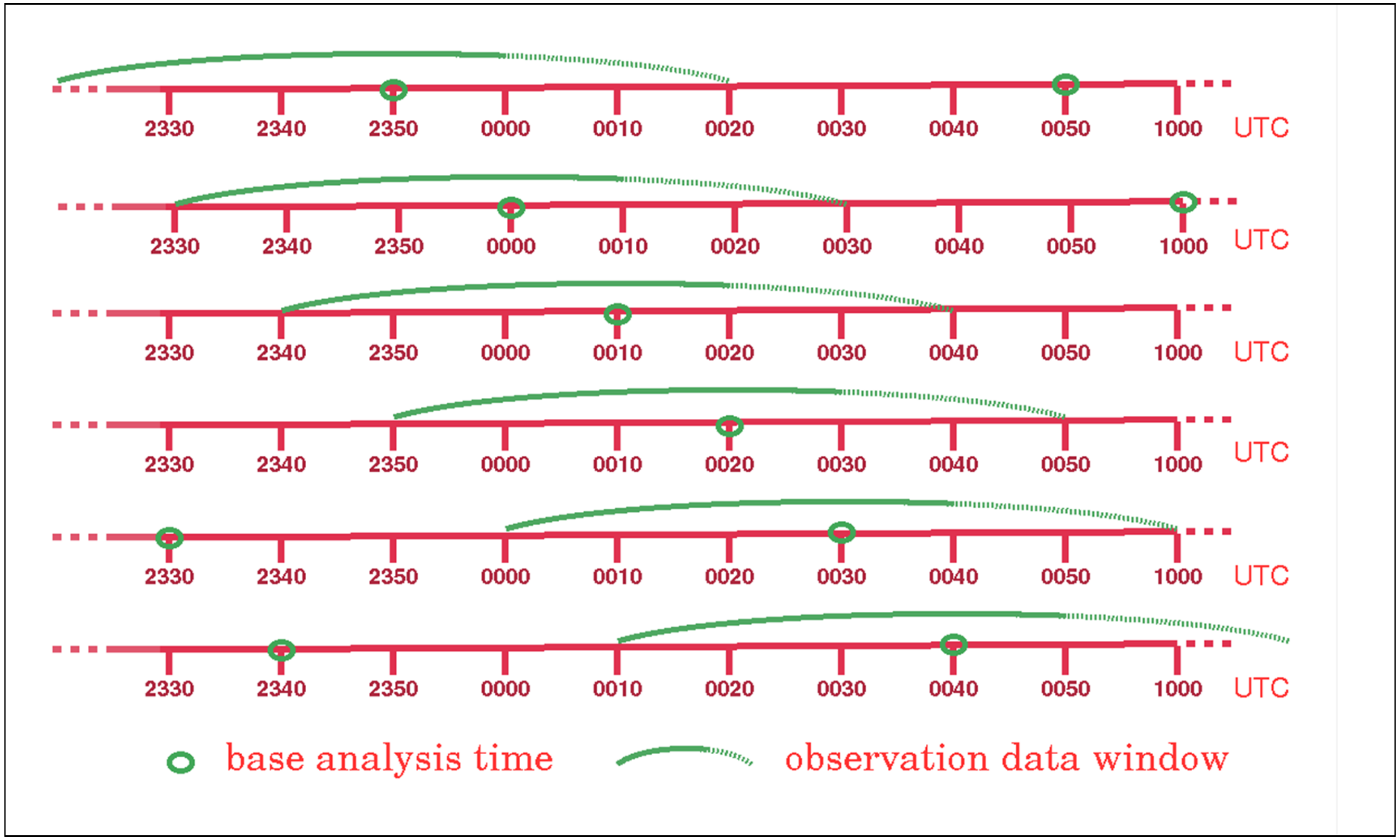


COMEPS = (3DVAR control on 3h window each hour: 4 perturbed members each hour) 6 hours = 18 perturbed members assembled each hour + 1 (control).

Improves usage of computing facilities.

Figure: Henrik Feddersen & Xiaohua Yang, DMI

DMI COMEPS Nowcasting product: Frequent Analysis with Overlapping Windows

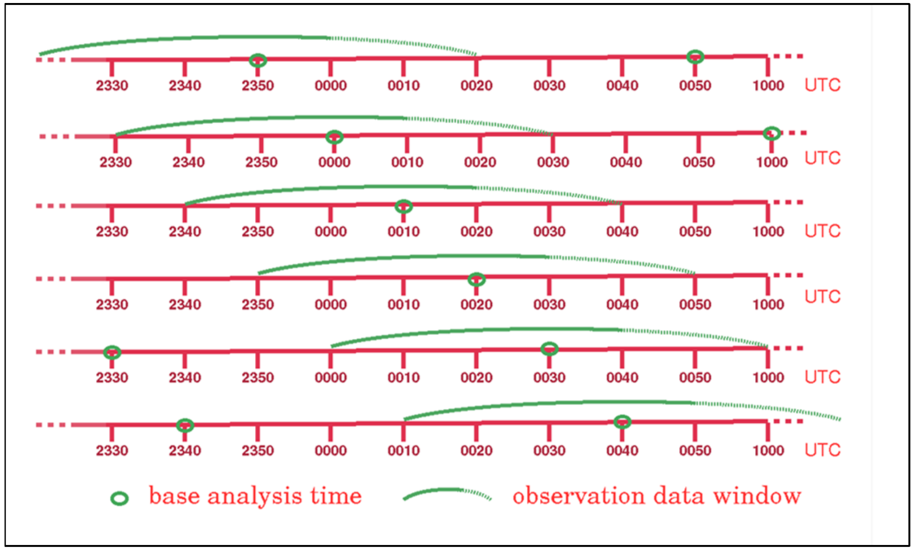


Harmonie-Lite (750m): new forecast every 10 minutes.

Each suite (rows) uses different observation batches in 3DVar.

SPOs (and other crowdsourced data) can enter one or more of these observation bathes.

Figure: Henrik Feddersen & Xiaohua Yang, DMI



Keep the frequent analysis cycle in mind

(same plot as previous slide)

Frequent Analysis with crowdsourced data

Can help improve observation usage in general (only $\approx$ 5-20 % of observations are used – crowdsourced is likely to be even less)

$\boldsymbol{R}$ is pre-set for each observation type, e.g., land-surface SYNOP pressure obs all have the same error.

The *Instrument error* is computed directly on the smartphone and is sent with the observation.

$$\boldsymbol{R}_{i,i} = \mathrm{R}^{\text{instrument}} + \mathrm{R}^{\text{representativeness}}$$

$$\mathrm{pdf}(\mathbf{x}|\mathbf{y}) \propto \exp\left(\underbrace{-\frac{1}{2}[(\mathbf{x}-\mathbf{x}_b)^T\mathbf{P}_b^{-1}(\mathbf{x}-\mathbf{x}_b) + (\mathbf{y}-H(\mathbf{x}))^T\mathbf{R}^{-1}(\mathbf{y}-H(\mathbf{x}))]}_{\text{Cost-function, } J(\mathbf{x})} \right)$$

Conclusions and recap

Motivation

- Crowdsourced data has both advantages and disadvantages over conventional data (e.g., low-cost observations, but poor quality).
- There exist many sources of data, only a few have been mentioned here.

Pressure from smartphones

- Data collection has been very successful. Screening works but can be improved, for example, with ML algorithms as shown by Conor McNicholas.
- SPOs have been successfully integrated into the HARMONIE setup

Remarks

- Third party data can be problematic because it is not always known how data have been processed before delivery.
- A unique user ID is used to bias-correct observations from each sensor. There are possible methods to solve this but not yet implemented.
- Legal issues must be considered, such as the GDPR act from the European Union. User consent is not the only way forward.