CA20108 Workshop on Gap Filling

COST Association Office, 23rd Floor

Avenue du Boulevard - Bolwerklaan 21, 1210 Brussels

August 29-30 2023

Participants: Amini, Setareh (Switzerland), Caluwaerts, Steven (Belgium), Dinh, Cuong (Ireland), Jacobs, Amber (Belgium), Koci, Ivan (Serbia), Lalic, Branislava (Serbia, Cost Action Chair), Louloudakis, Ioannis (Greece), Musyimi, Peter (Hungary), Nguyen, An (Ireland), Paschalidou, Anastasia (Greece), Roantree, Mark (Ireland, Workshop Chair), Slembrouck Joachim (Belgium), Stubner Ralf (Scientific Officer, Cost Association Office), Tanir Kayikci Emine (Turkey), Tasic, Visa (Serbia), Vergauwen Thomas (Belgium).

Workshop Presenters and Attendees

Session 1: Gap Analysis and Gap Filling

During this first session of the workshop, chaired by Mark Roantree, 3 scientists introduced their practices and ideas about gap filling and the importance of gap analysis.

Steven Caluwaerts (Ghent University & Royal Meteorological Institute of Belgium) presented his team's recent activities on gap analysis and gap filling. He first discussed the different reasons why his team needs gap filling. Firstly, if no solution is present to fill gaps, then the fraction of data that can be used for scientific analysis is strongly reduced. This is even more the case if different parameters (each with some gaps) need to be combined (e.g. to calculate heat stress indices) or if data from different stations are combined (e.g. a rural and urban station to calculate the urban heat island intensity). Additionally stakeholders and interested citizens do expect complete datasets for their applications. He expressed his surprise about the limited activity in the scientific community to address the gap filling challenge. Finally, he focused on the importance of gap analysis. Before you can decide upon a strategy to fill gaps, you first need to know your gaps (duration, frequency,…). Examples of gap distributions for Gent (see below) and Amsterdam datasets were presented.

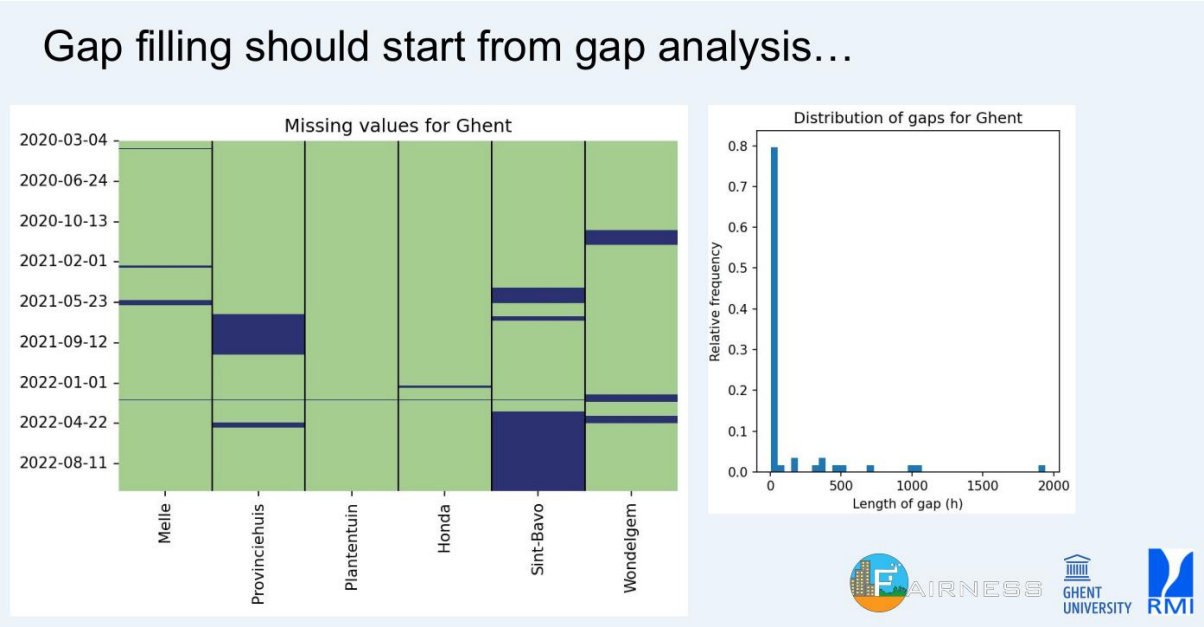

Figure 1: Gap Analysis

A toolkit called MetObs was presented by **Thomas Vergauwen** (Ghent University & Royal Meteorological Institute of Belgium). This Python package was recently developed for the FAIRNESS summer school in Gent and supports the processing of meteorological observations, from raw observations to analysis. It includes for example routines for synchronization, quality control but also gap filling. Regarding gap filling the ERA-5 debiasing based methods that will be presented by Amber Jacobs later during the workshop are included. The toolkit GUI was demonstrated by Thomas. The toolkit is well documented and freely available and very soon a publication will appear.

As final speaker **Ivan Koci** (PIS, Serbia), who is responsible for the PIS automatic weather stations (AWS) network in Serbia, shared his experience about gap filling. This network has collected a huge amount of data as it already exists for more than 10 years. Monitoring in the field conducted by PIS is based on monitoring of host plants, harmful organisms and environmental conditions. Environmental conditions are monitored using a network of automatic weather stations (AWS). Currently, there are 168 devices in the network that are placed deep in the green parts of the plants.

All of them produce hourly measurements of the following micrometeorological (µmet) elements: air temperature, relative air humidity, precipitation, and duration of leaf wetness. Some of the devices also produce measurements of soil temperature, solar radiation, wind speed, and direction. The devices are organized into clusters according to the manufacturers and their rules for the transfer of measured data and storage in databases. Measured data are downloaded from cluster databases and placed in the main µmet database for processing and production of various biometeorological products (Figure 2). The preferred data access is by REST APIs (where possible, where not there are DLLs). It is important to emphasize here that the network of AWS, organization, and access to measurements and processing of measured µmet data has a primarily operational character through pairing with observed biological data at the same locations.

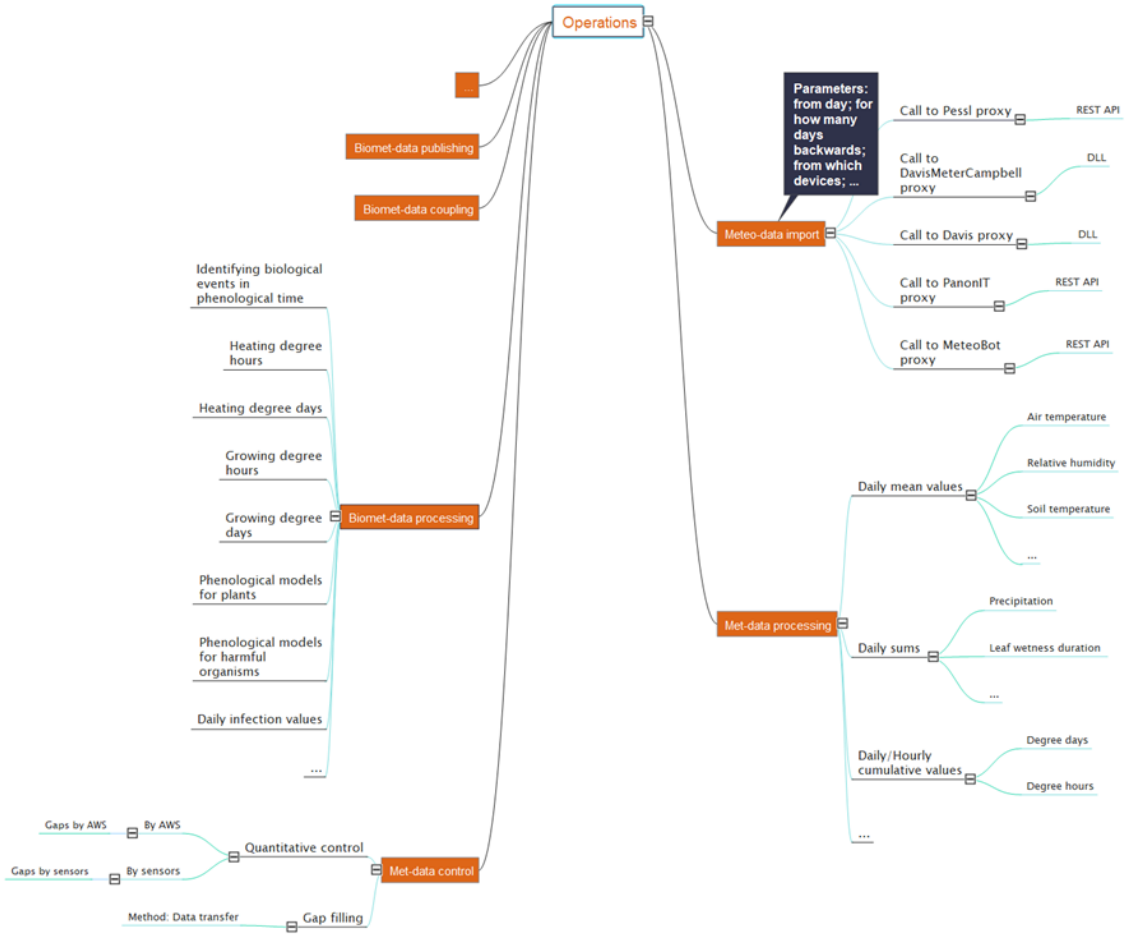

Figure 2: Operational character of PIS network and µmet data management: activities and tasks

The data are used on a daily basis by agricultural stakeholders to make decisions about for example spraying. It is therefore utterly important to provide reliable information. Using an interactive dashboard (Figure 3) he gave insights in the typical gaps he is confronted with. For the moment there was no time to develop an automated gap filling or quality control procedure but Ivan explained that such a system would be very useful for their network.

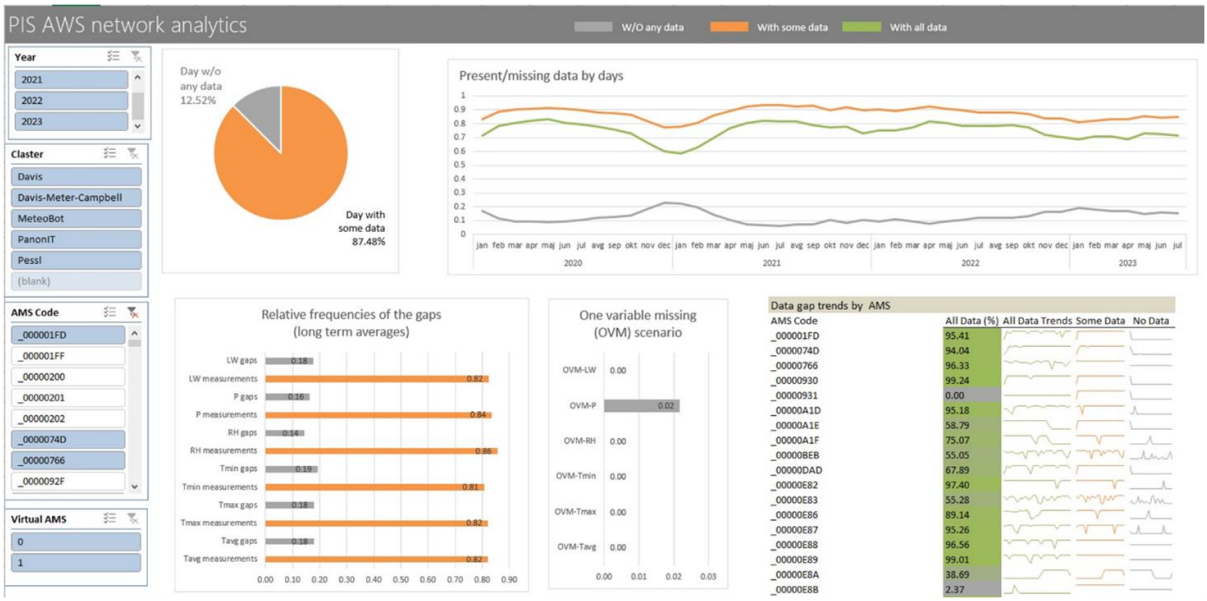

Figure 3: PIS network - hourly gaps distribution dashboard

Session 2: Gap Filling with Python

The second session, chaired by Branislava Lalic, focused on gap filling using Python's *scipy* library of functions. The team from Dublin City University (Mark Roantree, Dinh Viet Cuong, An Nguyen) presented their work on approaches to gap filling with Python. The presentation focused on Python's built-in functions to ensure minimal expertise in terms of programming or machine learning. Three 1-dimensional gap filling methods were introduced and discussed: Linear, Nearest and Spline. In addition, three n-dimensional methods were presented: Linear, Nearest and the Radial Base function. Interpolation using Python functions works by first creating a plot that is the closest fit to the target dataset. Once this plot has been determined, the interpolation process is relatively straightforward.

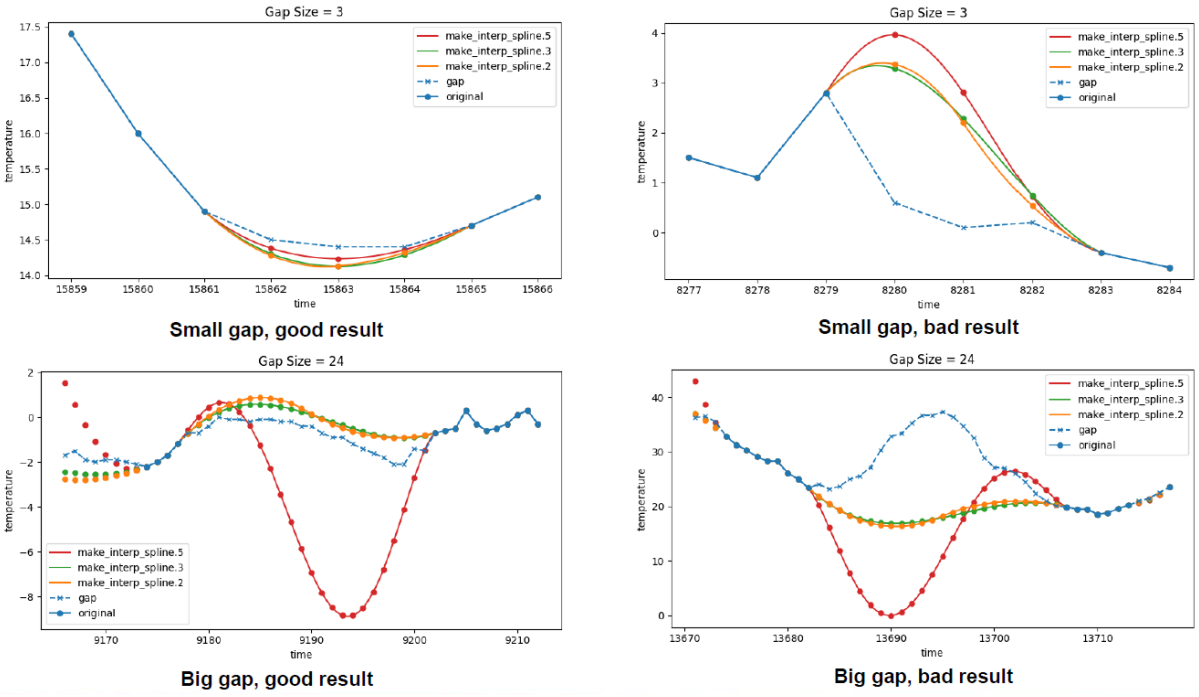

Figure 4: Spline Interpolation & Analysis of Results

One-Dimensional Gap Filling Methods.

From the *scipy.interpolate* library, the interp1d function was used for both Nearest and Linear; the *make_interp_spline* function was used for Splines; the more advanced *Akima1DInterpolator* function was used as a piecewise function using cubic splines to plot data; and the *PchipInterpolator* function another form of cubic interpolation.

Figure 4 shows the different results obtained during experiments with the Nearest interpolation function. In summary, it is possible to obtain bad results even for small gaps (this function does not create a good plot for this dataset) while on the other hand, very good results can be obtained for small gaps while relatively good results were obtained for large gap sizes. This section of the presentation finished with a comparison of 4 functions using RMSE scores.

N-Dimensional Gap Filling Methods.

In terms of multi-dimensional gap filling, the scipy.interpolate library provides the *NearestNDInterpolator* function for Nearest interpolations; the *LinearNDInterpolator* for Linear interpolations; and the *RBFInterpolator* function for more powerful interpolations using Radial Basis Functions. Figure 5 shows the types of plot fitting capabilities using 2 forms of RBF interpolation.

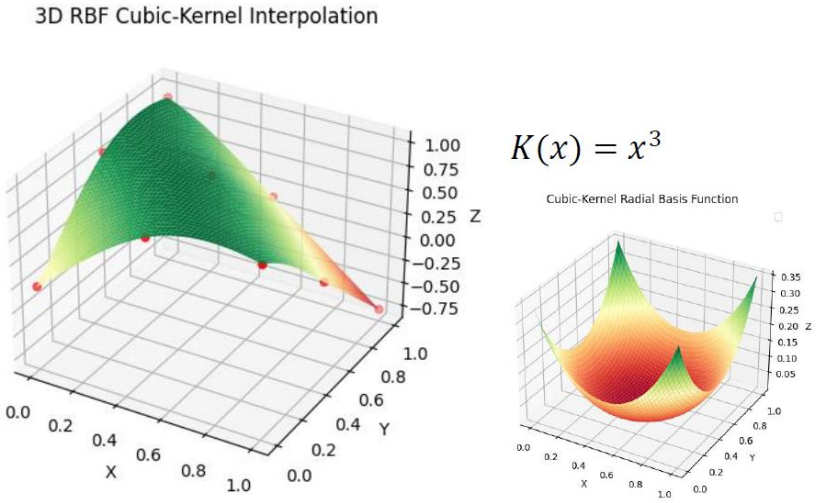

Figure 5: Curve Fitting Using 3-D Functions

As with the 1-Dimensional functions, a comparison across the different N-Dimensional functions was presented. This is shown in figure 6 for where the RBF interpolation showed the strongest set of overall results for the dataset used in the evaluation.

N-D Gap Filling- Comparing Results

RMSE Increase with Gap Size (but …..)

Method \ Gap Size	1	3	6	12	24	48
NearestNDInterpolator	0.9294	0.9564	0.9711	0.9641	0.9627	0.9566
LinearNDInterpolator	0.6907	0.9031	1.3067	2.2631	3.2734	3.4431
RBFInterpolator.linear	0.5910	0.7240	0.7980	0.8379	0.8518	0.8587
RBFInterpolator.thin_plate_spline	0.5465	0.7456	0.8659	0.9622	1.0069	1.0284
RBFInterpolator.cubic	0.5179	0.7987	1.0257	1.2653	1.3829	1.4447
RBFInterpolator.gaussian	0.9062	1.6513	1.8385	1.9440	1.9541	1.9455

Figure 6: Comparing Results Across Gap Sizes for 6 different 3-D Interpolation Functions.

Session 3. Traditional Approaches to Gap Filling

The third session on traditional approaches to gap filling was chaired by Steven Caluwaerts. Branislava Lalic (University of Novi Sad, Serbia) presented an overview of gap-filling techniques for different durations of gaps. The methodology presented by Lompar et al. (2019) is developed based on the debias of ERA5 reanalysis data. The methodology is tested for different landscapes, latitudes, and altitudes, including tropical and midlatitudes (Table 1).

Table 1 Locations used in the study

Location	Time series	Landscape	Position			ERA5		
			Lat. (°)	Long. (°)	Alt. (m)	Lat. (°)	Long. (°)	Alt. (m)
Kikinda	2014-2017	Lowland	45.87	20.46	82	45.9	20.4	73
Gumpenstein	2014-2017	Mountains	47.49	14.09	700	47.4	14.1	1080
Bahariya	2017	Desert	28.41	28.93	99	28.4	28.9	97
Montecristo	2016	Island	42.34	10.31	645	42.3	10.2	645
Pianosa	2016	Island	42.58	10.08	29	42.6	10.2	29

The debias procedure included hourly measured air temperature. For a 30-day learning period before the missing observation (Figure 7a), the relationship between the existing observations and reanalysis is determined with that relationship then applied to the reanalysis value to calculate the missing observation. In addition, linear regression was used to determine the relationship between observations and reanalysis for the learning period. The length of the learning period was chosen empirically (short enough to be influenced by seasonal changes but long enough to contain an appropriate amount of data for linear regression), and linear regression was chosen because it is a computationally efficient and fast process that satisfies the requirements of this methodology. Due to uncertainties in radiation, surface, and PBL physics schemes in numerical models, diurnal variations in temperature biases can occur. To avoid this problem, a ± 3-hour filter was applied to the time series. In this way, data from the same part of the day were used. After the removal of diurnal changes in bias with the filter on the learning dataset, the linear regression coefficients were calculated (Figure 7b) and applied to the reanalysis values for missing observation dates and times.

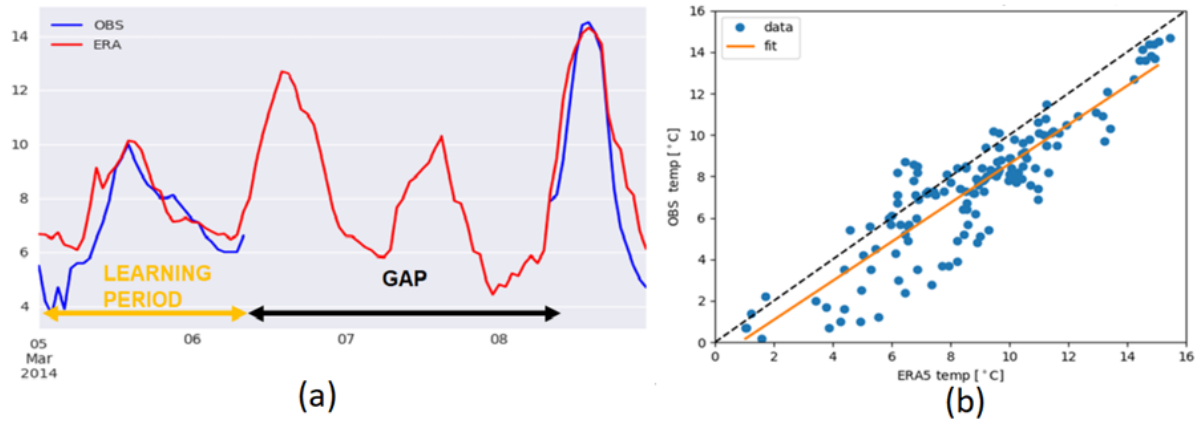

Figure 7. (a) Time series with a gap in temperature observations. The blue line represents observations, and the red line represents ERA5 values for the nearest point. (b) Linear regression of learning data for one time step in the gap.

The efficacy of the used gap-filling technique is presented in Table 2. At all locations and for all gap scales, the maximum $RMSE_{DEB}$ calculated using daily data is smaller than the results obtained using hourly data. However, differences between $RMSE_{DEB}$ and $RMSE_{ERA5}$ are less pronounced in the case of daily data except in the case of Gumpenstein where the maximum deviation for daily data is highest. Almost the same results for ($RMSE_{DEB}$ - $RMSE_{ERA5}$) obtained using hourly and daily gap-filling data clearly indicate that there is a systematic deviation of ERA5 data from measurements.

Table 2 Maximum absolute values of ($RMSE_{DEB}$ - $RMSE_{ERA5}$) difference and $RMSE_{DEB}$ calculated using hourly and daily data for gap filling.

Location	Bahariya (°C)	Gumpenstein (°C)	Kikinda (°C)	Montecristo (°C)	Pianosa (°C)
max(|$RMSE_{DEB}$-$RMSE_{ERA5}$|) for hourly data	1.81	6.34	0.87	0.95	2.12
max(|$RMSE_{DEB}$-$RMSE_{ERA5}$|) for daily average	1.31	7.32	1.03	1.08	1.63
max(|$RMSE_{DEB}$|) for hourly data	2.54	4.19	2.21	2.18	2.31
max(|$RMSE_{DEB}$|) for daily average	1.82	4.02	1.69	2.06	1.98

Additionally, it is important to address the fact that gap-filling techniques are overwhelmingly dealing with air temperature time series. Analysis of long-term data series (Figure 3) witnesses gaps commonly appearing due to AWS work failure, leaving gaps in data series of all measured data, typically air relative humidity and precipitation. Regarding relative humidity, it is reasonable to expect a similar efficacy of the presented methodology. In case of precipitation gaps, it is not the case. Therefore, filling gaps in precipitation data series requests an entirely new approach.

For the final presentation, Gap-filling urban observational data using ERA-5 debiasing was presented by Amber Jacobs from the University of Ghent. This was a presentation of Amber's development of a novel debiasing technique for ERA-5 data and is due to be submitted for publication shortly.

Session 4. Topics for Future Research

The final session, chaired by Mark Roantree, focused on potential collaborations between the researchers at the workshop and wider collaboration across the cost action. A number of possibilities were discussed including comparing the performance of the MetObs toolkit (U Ghent) with the Python interpolation function (DCU). Different datasets with different characteristics/gap distributions were presented:

- Ivan Koci – PIS dataset (target variable temperature). QC data first then use as validation dataset for 2-method project. 100% data available for validation
- Visa Tasic AWS data (target variable temperature). (wind direction). No 100% dataset so validation will require a revised method.
- Setareh Amini. QC effort for Homogenisation of data. Temperature gaps require filling.

In addition, a collaboration under the topic of **Climate Data and Health with** Anastasia Paschalidou using Epidemiological and temperature data was discussed. This could involve researchers from Ireland and Greece. A third project was discussed, based on **Evapotranspiration in Kenya with** Peter Musyimi using Soil Moisture prediction from Temp, Humidity, Dewpoint which could potentially involve researchers from the Ghent team.

References.

1. Lompar M, Lalić B, Dekić Lj, Petrić M (2019) Filling gaps in hourly air temperature data using debiased ERA5 data, *Atmosphere10* (1), 13; https://doi.org/10.3390/atmos10010013