RINGO | Readiness of ICOS

Readiness of ICOS for Necessities of integrated Global Observations

D4.3

Reprocessed long term data series from 9 Ecosystem stations

RINGO (GA no 730944)

Public Document/define level

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730944

Deliverable: Concept on standards for data collection, classification, description, processing and distribution and methods for data identification, traceability and sharing in FLUXNET

Author(s): Dario Papale, Simone Sabbatini, Sundas Shaukat, station teams, Carbon Portal

Date: December 7th 2020

Activity: WP4 Task2

Lead Partner: ICOS ERIC

Document Issue:

Dissemination Level: Public

Contact: darpap@unitus.it

	Name	Partner	Date
From	Dario Papale	UNITUS	2/12/2020
Reviewed by	Alex Vermeulen	ICOS ERIC	8/12/2020
Approved by	Elena Saltikoff	ICOS ERIC	29/12/2020

Version	Date	Comments/Changes	Author/Partner

Deliverable Review Checklist

A list of checkpoints has been created to be ticked off by the Task Leader before finalizing the deliverable. These checkpoints are incorporated into the deliverable template where the Task Leader must tick off the list.

	√
• Appearance is generally appealing and according to the RINGO template. Cover page has been updated according to the Deliverable details.	☐
• The executive summary is provided giving a short and to the point description of the deliverable.	☐
• All abbreviations are explained in a separate list.	☐
• All references are listed in a concise list.	☐
• The deliverable clearly identifies all contributions from partners and justifies the resources used.	☐
• A full spell check has been executed and is completed.	☐

DISCLAIMER

This document has been produced in the context of the *project* Readiness of ICOS for Necessities of integrated Global Observations (RINGO)

The Research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730944. All Information in this document is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability. For the avoidance of all doubts, the European Commission has no liability in respect of this document, which is merely representing the authors view.

Amendments, comments and suggestions should be sent to the authors.

EXECUTIVE SUMMARY

In this report the summary of the legacy data production is reported. The data are loaded in the RINGO Fileshare together with the informastion needed to interpret them.

The exercise helped to highlight a number of important aspects:

1) It is only possible to process legacy raw data centrally at the desired quality if all relevant metadata are provided
2) The use of different formats, especially if not well documented, turns the processing more complex and labor intensive
3) It is crucial to correctly collect, organize and store all the data and related metadata in all detail, to avoid the risk that the measurement data become unusable.
4) The comparison between reprocessed and historical version of the legacy fluxes shows a general good agreement in terms of the interannual variability pattern, however some exceptions and bias probably due to a different choise in specific corrections and the unavailability of important metadata have been identified
5) It is recommended to add the timeseries of the legacy data of Class1 and Class2 stations as separate ICOS product if the standard of the Associated station is followed, where the important metadata are collected and shared. This could also include, if needed, a reprocessing starting from the raw data
6) It is important to prepare a metadata set to describe in details the raw data processing in a machine readable format (FAIR) and this is a task that the ICOS ETC should focus on in collaboration with the other networks of eddy covariance sites.

1 Introduction

In recent years there has been a remarkable increase in the number of eddy covariance (EC) monitoring stations worldwide, all using different sonic anemometers and gas analyzers, based on different technologies and characteristics, to measure the vertical exchange fluxes of gases, energy and water. Use of different sensors at different EC stations and complex processing steps introduce uncertainties and different ranges of variability in the calculated fluxes. In ICOS Ecosystem it has been decided to use a standard system, composed by the same type of sensors among stations and with a setup made according to specific instructions (Sabbatini and Papale 2017). In addition, the raw data collected at the ICOS stations are processed centrally at the Ecosystem Thematic Centre (ETC), reducing the cross-sites variability due to choices in the processing methods, implementation and data processing pipeline.

Most of the ICOS Ecosystem stations are however sites where for a long period before ICOS was established (in some cases more than 20) already eddy covariance data have been collected and at that time processed using different sensors, setups and processing schemes/software. These legacy data are unique and of great value and have been used in many studies and shared in a number of topical databases. With any change of setup and processing there is the risk to introduce (or better remove) biases and other differences that should be at least quantified and if possible minimized. In addition, the past data are not in line with the ICOS standard in terms of completeness of metadata, raw data collected and processing and a strategy to reconcile the different available datasets should be discussed and implemented.

The objective of the activity in the context of WP4, Task4.2, is to make these historical timeseries of calculated fluxes available using a processing routine that is as much as possible complaint with the one applied in the ICOS processing. In this way we will approximate the needed continuity in the EC timeseries.

Raw data and all the available metadata from these sites have been collected and the data has been processed in the ETC using the standard methods. The final aim was not only the provision of the dataset but also an analysis of the critical aspects, a quantification of the resources needed and a strategy to integrate this additional dataset in the context of the ICOS data system at the Carbon Portal.

2 Data collection and processing

Nine sites have been selected during the project proposal preparation to provide legacy data for the test (see Table 1). In some sites the legacy data at one point uses the ICOS setup, that is always composed by a Gill HS ultrasonic anemometer and a LI-COR 7200 Infrared Gas Analyser. The standard ETC processing applied is described in Sabbatini et al. 2018 and includes four different processing options that lead to an uncertainty range definition. The code used is also available in the ETC GitHub portal.

Table 1: Details of the stations involved in the analysis. All the sites have currently the ICOS setup operational (LI-7200 and Gill HS). For 7 stations there has been a period of overlap between the historical and the new setup, and results of this parallel measurements are in the Milestone MS30. Period and setup of the legacy data is reported. CRO=cropland, EBF= evergreen broadleaf forest, ENF= evergreen needle leaf forest, WET= wetland.

Stations		Legacy data	
Sites	**Ecosystems**	**Years**	**Setups**
CZ-BK1	ENF	13	Gill R3, LI-7000
CZ-wet	WET	12	Gill R3, LI-7500
DE-Tha	ENF	17	Gill R3, LI-7000
DE-Geb	CRO	16	Gill R3, LI-7000
FR-Bil	ENF	5	Gill R3, LI-7500
FR-Pue	EBF	14	Gill R3, LI-6262
BE-Lon	CRO	9	Gill R3, LI-7000
FI-Hyy	ENF	23	Gill R2, LI-6262
FI-Sii	WET	13	Metek, LI-7000

Seven out of nine sites had also the historical setup still in place for a period in parallel with the ICOS setup, allowing a specific analysis of the differences that has been presented and included in the Milestone 30 report. Here the summary and introduction to the legacy data is presented together with the analysis of the critical aspects to consider in case it is decided to proceed with the reprocessing for all the ICOS stations that have a legacy dataset.

3 Results

For all the nine sites the raw data have been collected together with the metadata. For each site a number of subgroups have been created according to 1) the format of the raw data and 2) possible different setups. Raw data have been loaded in the FileShare repository of the RINGO project:

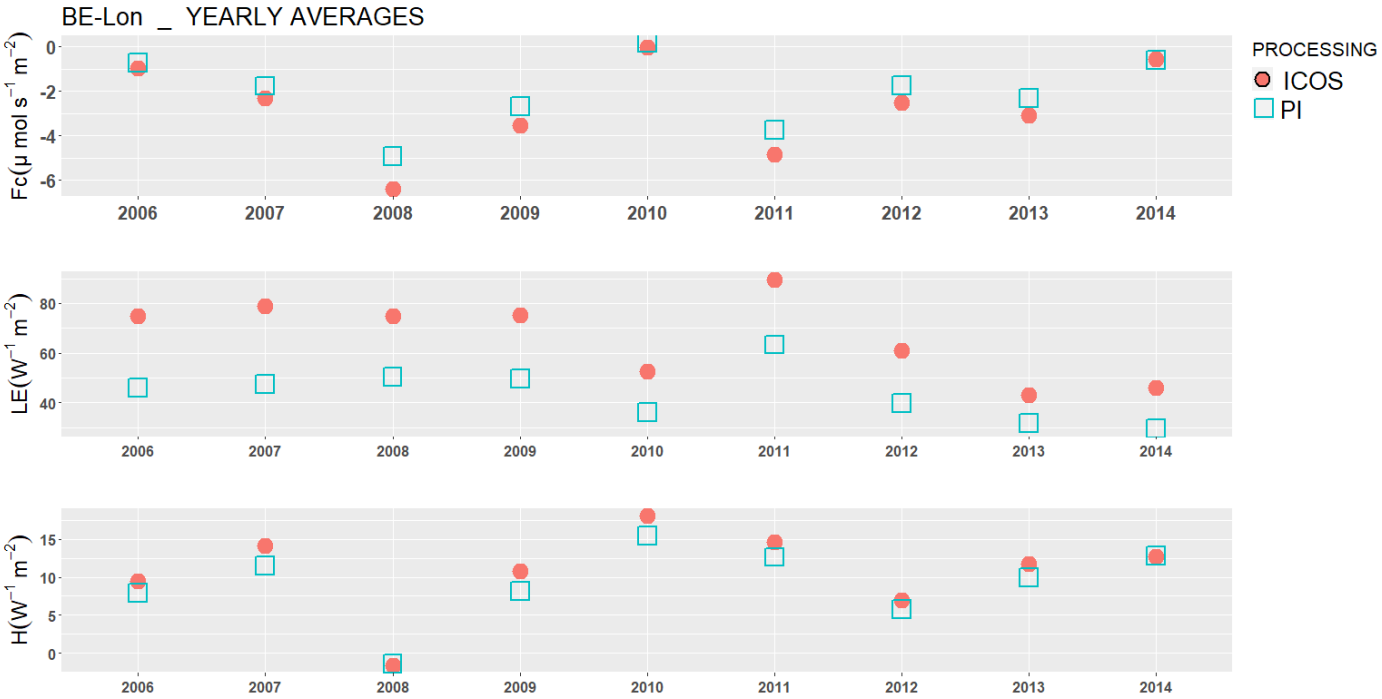
https://fileshare.icos-cp.eu/s/XBa2ftDaoJfXfyM

The total amount of storage capacity needed for the raw data for the 9 sites was about 2 Tb.

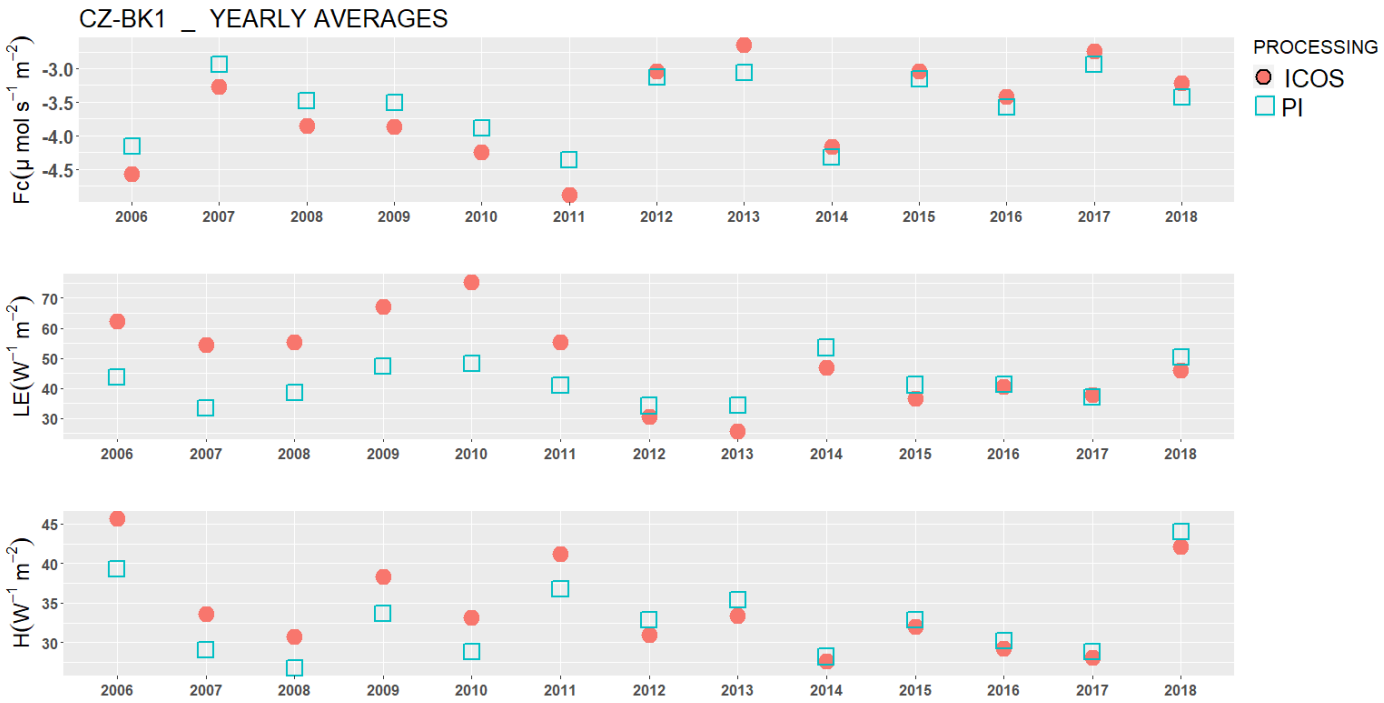
The processing has been carried out using the ICOS Carbon Portal parallel computing facilities with a remote control of the process. The results, all with a same standard format, are available in the RINGO groupdisk in FileShare at the address:

https://fileshare.icos-cp.eu/s/xmRq2xDrSgg6RZ6

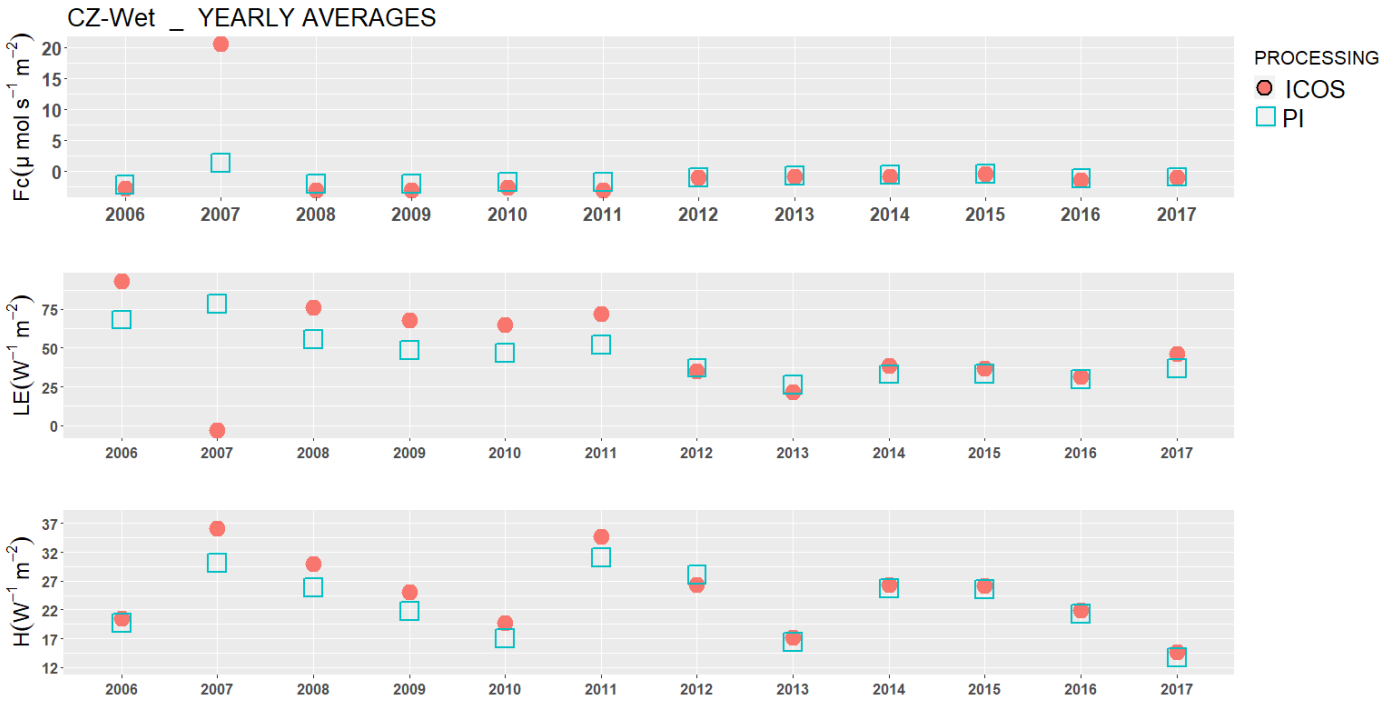
In order to have an overview of the general variability and agreement between the centralized processing and the PI versions, the legacy data results obtained by the ETC have been compared with the official version uploaded by the station team and available in the European Flux Database (www.europe-fluxdata.eu) or in the FLUXNET2015 collection (Pastorello et al. 2020). The following plots show the annual averages of the two versions of the legacy data (ICOS version processed by the ETC and PI version, the one officially available in the databases.



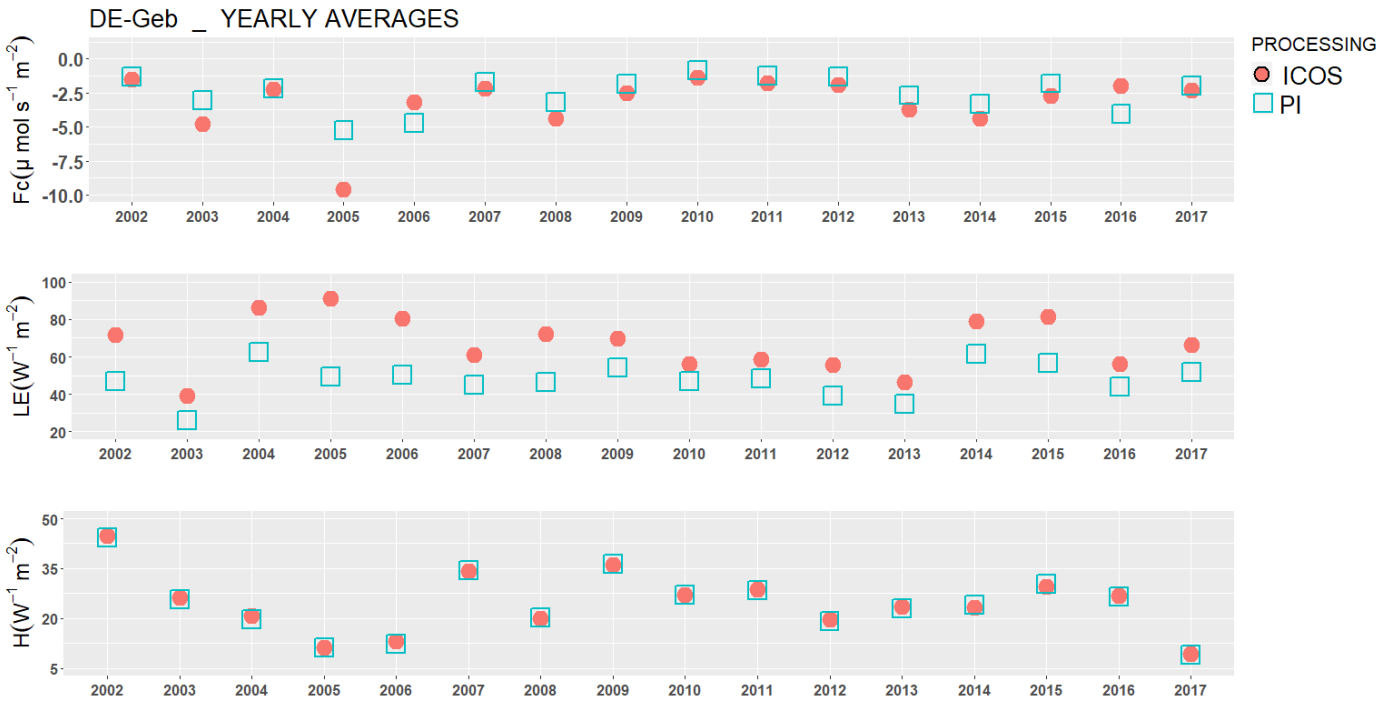
Note: in 2014 the ICOS setup has been implemented. Close path IRGA with 12m tube. PI processing using Eddysoft



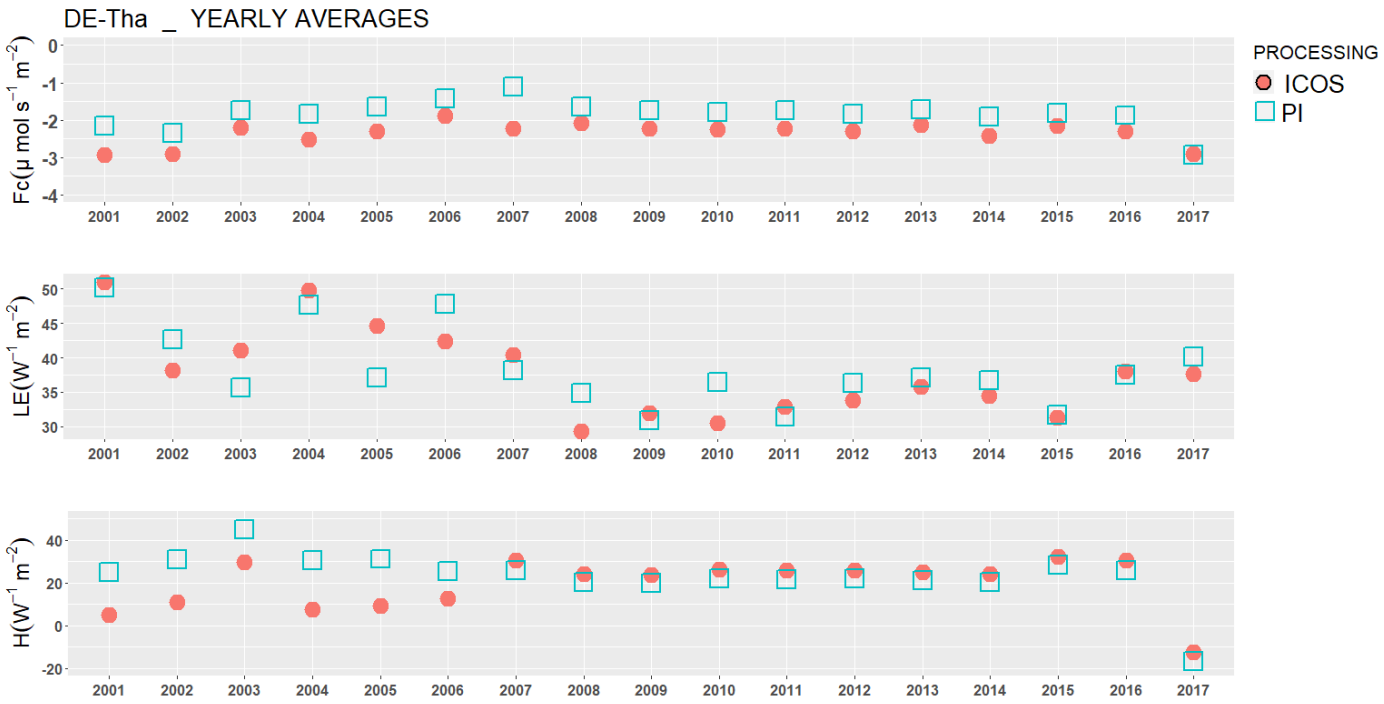
Note: in 2011 the ICOS setup has been implemented. Close path IRGA with 5m tube. PI processing using Eddypro



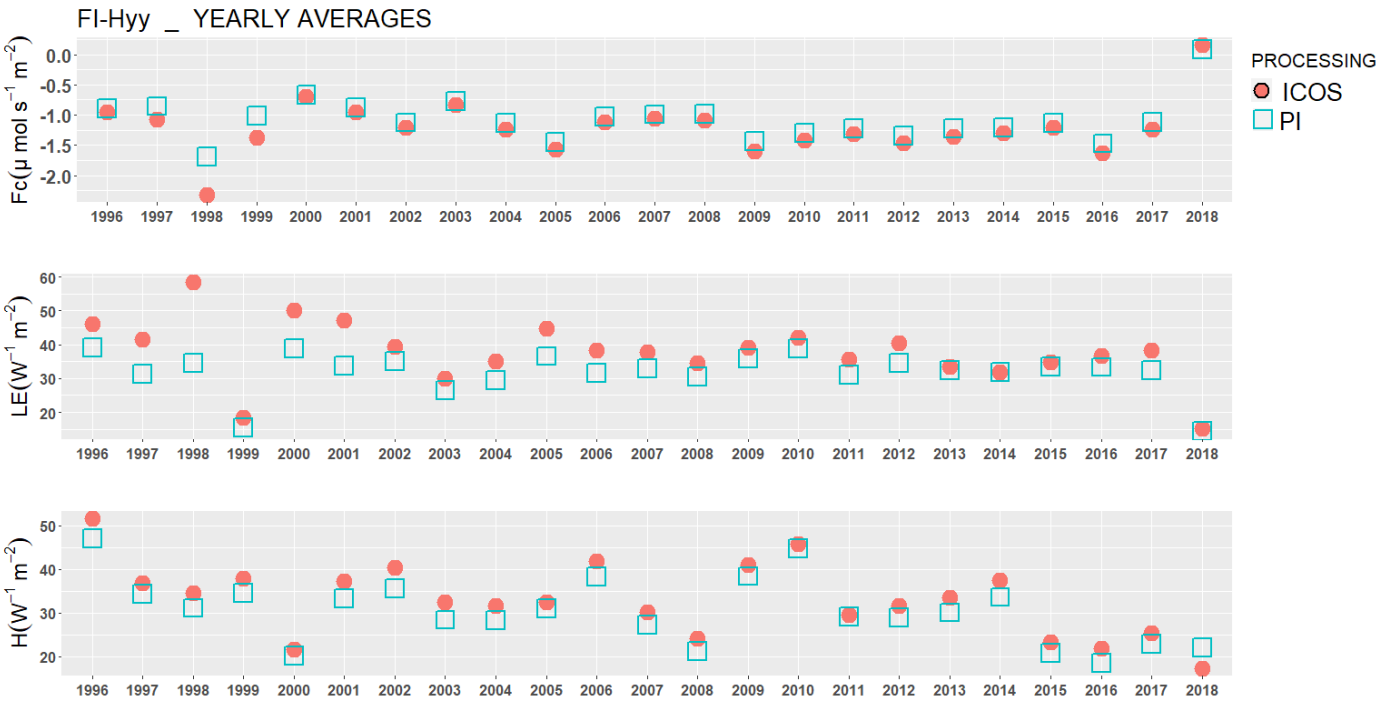
Note: in 2011 the ICOS setup has been implemented. PI processing using Edire



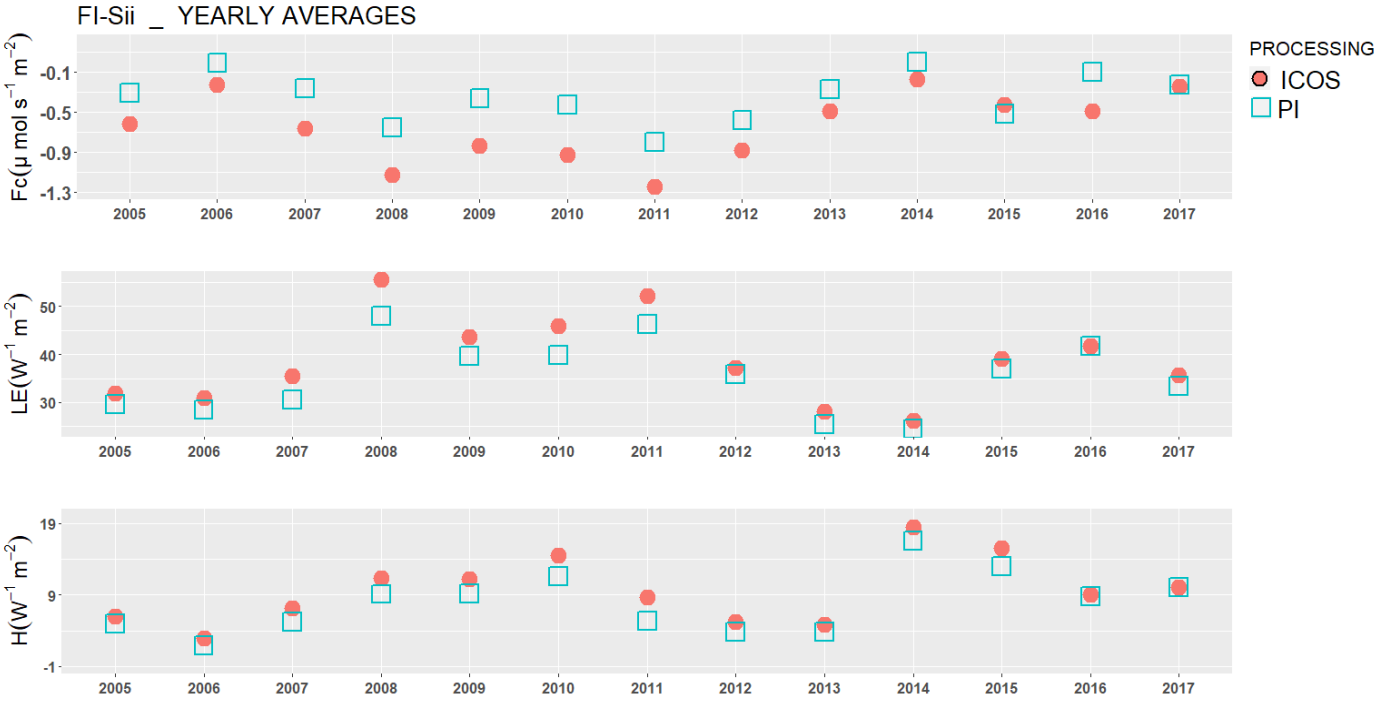
Note: in 2017 the ICOS setup has been implemented. Close path IRGA with 12m tube. PI processing using Eddypro



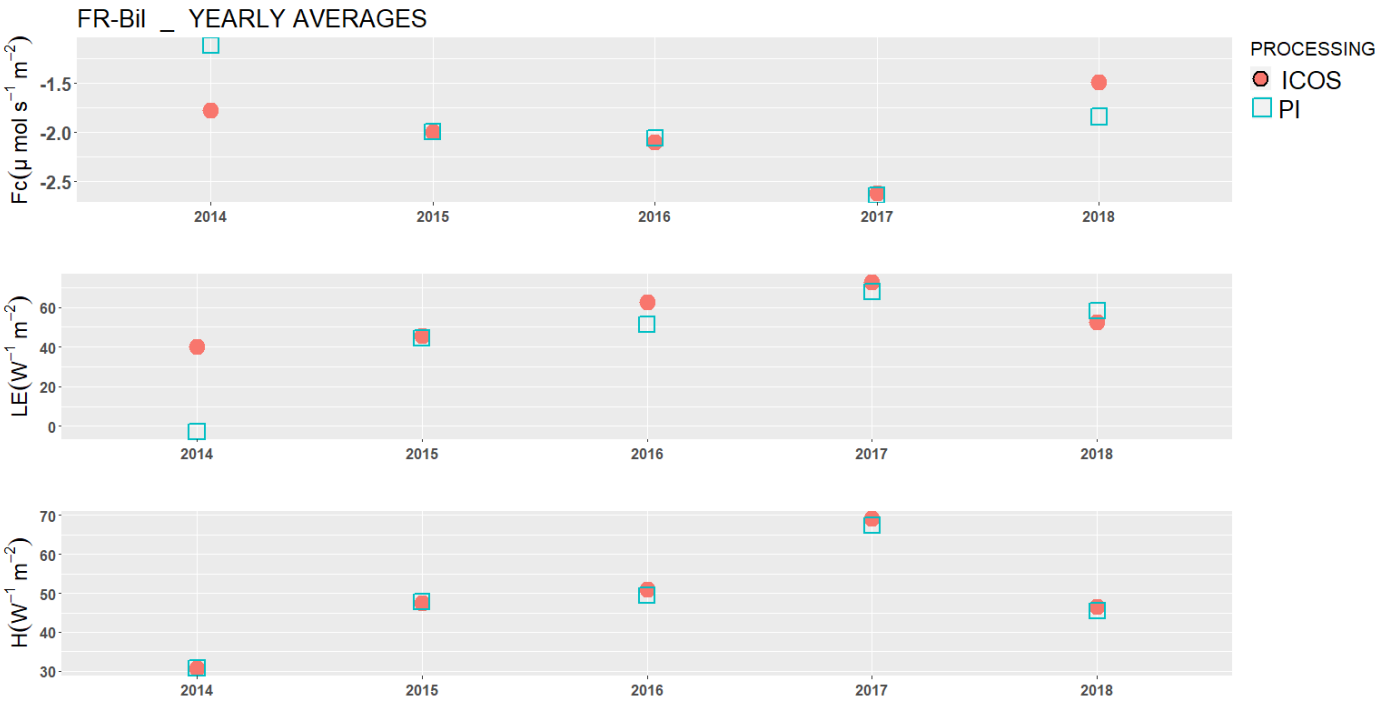
Note: in 2017 the ICOS setup has been implemented. Close path IRGA with 63m tube. PI processing using Eddypro



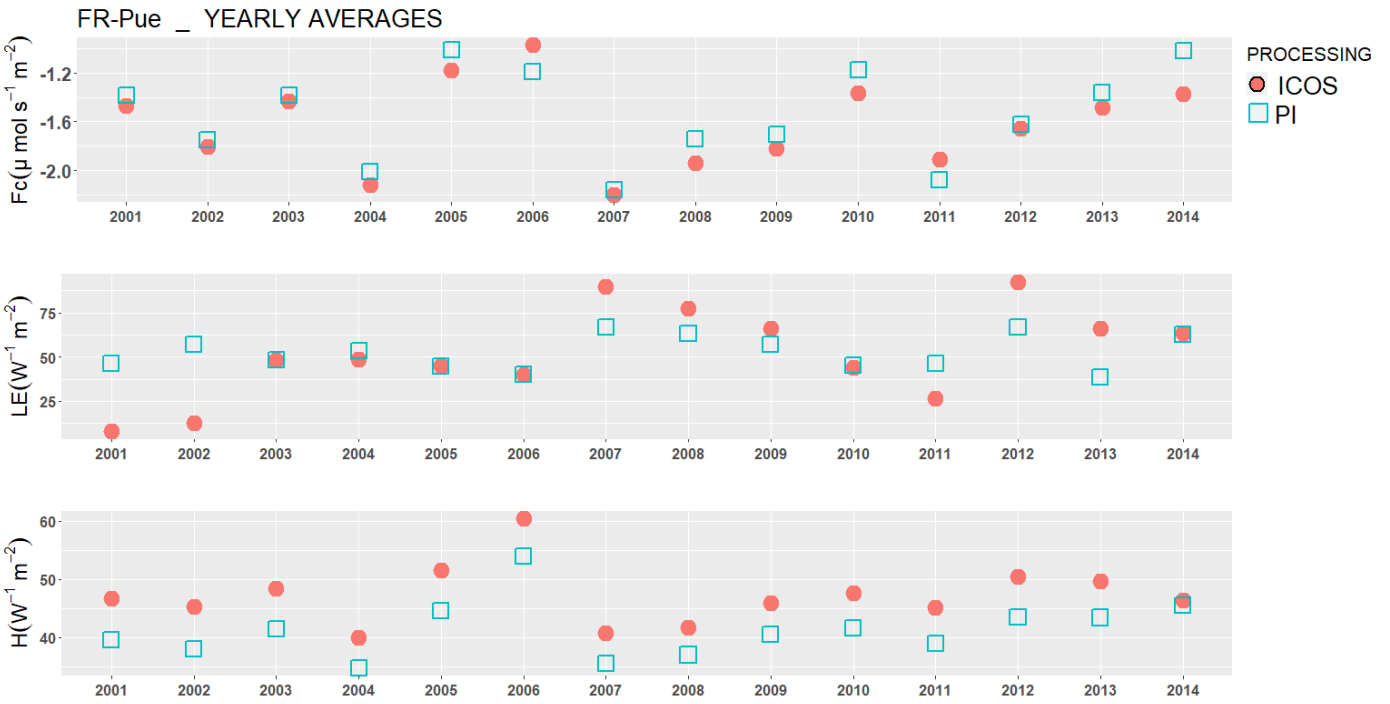
Note: in 2018 the ICOS setup has been implemented. Close path IRGA with 8m tube. PI processing using EddyUH



Note: in 2015 the ICOS setup has been implemented. Close path IRGA with 13m tube. PI processing using EddyUH



Note: in 2015 the ICOS setup has been implemented. PI processing using Eddypro



Note: in 2014 the ICOS setup has been implemented. Close path IRGA with 3.5m tube. PI processing using Eddypro

A detailed analysis of the differences and explanation for discrepancies between the ICOS ETC version and the Station Team version is out of the scope of this RINGO Task and would require an intensive and detailed analysis that is impossible give the resources available. The data availability (raw and processed) together with the code used by the ETC makes however the analysis possible in future or at level of single group.

In general however it is possible to note that:

1) There is a general agreement in the interannual pattern between the two versions, although there are years with larger differences and there are different cases with an offset between the two.

2) Considering only the years where the ICOS setup was installed the agreement is higher, this also because the processing in the ETC is optimized for this setup.
3) While the two processing applied the same coordinate rotation and detrending methods (the one communicated by the station teams), the spectral correction methods were different and heterogeneous and could be a reason for the biases.
4) In the Latent Heat fluxes in particular, the tube effect is important when closed path systems were used (LI7000) and the spectral correction used can have a large impact (in ICOS ETC setup the Fratini et al. 2012 was used and this takes the effect into account).
5) We can be sure about the fact that the processing applied centrally in the ICOS ETC is the same across the years while this could be not ensured for the PI version where changes are possible between years.

4 Quantification of the effort and main issues

The exercise allowed the quantification of the effort for the legacy data processing in terms of PM (from the data collection and organization to the calculation that was done in this case in the CP cluster). This is clearly function of how well the data are organized at the site, how rich are the metadata information available and how much the raw data are standardized. For this reason, it is difficult to give a general estimation for the first part (data collection and transmission) that can vary between 0.1 and 1 PM.

Also, the centralized data preparation is function of the original data organization and number of different formats available. On average only for the identification of the different formats, their definition in the import routine for the processing and the preparation of the setup files needed to run the processing in parallel took about 0.3-0.5 PM per site. The processing and the analysis of the results to ensure that all worked as expected required about 1-1.5PM in total.

We also identified in general terms the main difficulties and issues that should be carefully considered in the data acquisition and organization to allow future reprocessing and data interpretation. All the issues listed here below were found to be critical for at least one of the nine sites processed:

1) Raw data format: the raw data should be possibly in the same format for the same site (across years) to simplify the reprocessing. In case this is not possible it would be important to have standard and clear descriptions of the different formats and the periods covered.
2) The metadata describing the measurement setup (sensor models and SN, height of measurement, tube characteristics, pup and flow rate, configuration, alignment and orientation of the sonic etc.) should be recorded and safely stored, if possible in standard and machine readable formats.
3) The metadata and information about all the details related to the data processing (calculation methods, corrections applied, quality filtering scheme used etc.) and the software or code used should be available in machine readable format, and possibly also the code used shared open source.
4) All the parameters used for the conversion from instrument units to physical units should be safely stored (there have been cases where it has been impossible to convert mV to physical quantities)
5) Old data are often collected using analogue signals and for this reason without any sensor diagnostic, while these are crucial to identify malfunctioning periods in the timeseries.
6) All the data and metadata should be safely stored in repositories that will make them available to the station team. This is particularly relevant in long timeseries where different PIs and also different postdoc, PhD and technicians work at the same site during the years.

This list is clearly not exhaustive but gives an idea of the main issues and for this reason helps top evaluate what is currently done in ICOS and where things should be improved. The high level of standardization of setup, data format and metadata system ensure that issues number 1, 2, 4 and 6 are well covered, also thanks to the robust system in place at the Carbon Portal in terms of data repositories and backups. The digital acquisition decided in ICOS and the list of variables collected, that includes also a large number of additional parameters and diagnostic values produced by the sensors, makes the ICOS data also compliant respect to the issue number 5.

Where improvements can and should be done is on the issue number 3. In fact, although the method used is documented in open access peer reviewed papers and the processing code is available in the ICOS ETC GitHub system, a machine readable (FAIR) list of all the processing steps applied and relative parameters is still missing. This is for sure an important improvement where the ICOS ecosystem community and in particular the ICOS ETC have to invest resources in the near future, in agreement with the other international networks.

5 How to make the legacy data available

The ICOS Ecosystem network is composed of two categories of station, the Class 1 and 2 that are highly standardized and where the raw data are collected and processed, and the Associated station where the data shared are half hourly fluxes processed by the station teams and collected using different setups and sensors (Table 2). For this reason the Associated stations can submit also data collected before the labelling date if they follow the requirements of quality and documentation. The legacy data of ICOS stations (both Class 1 and 2 and Associated) are instead shared in other databases (European Database, FLUXNET etc.) following a more basic and minimal standard requirement.

Table 2: characteristics of the dataset submitted and shared by the different type of stations.

	ICOS Class 1 and 2	ICOS Associated	Legacy (today)
Standard setup	YES	NO	NO
Centralized flux calculation	YES	NO	NO
Per-sensor variables*	YES	YES	NO
Full setup metadata	YES	YES	NO
Processing metadata	(YES)*	(YES)*	NO

** the variable measured by each single sensor are submitted individually, without any spatial aggregation that is instead done in the post-processing. This ensure better quality and possibility to estimate spatial variability indicators in a standard way. (YES) indicates an information planned but not yet fully operational.*

The Associated stations can, according to the scheme reported in Table 2, prepare the legacy data following the requirements and have the full legacy timeseries shared in the context of ICOS and with the same quality level.

In case of Class 1 and Class 2 stations this is currently not possible, because the quality of the data and the compatibility with the Level 2 data produced by the ICOS ETC (in terms of standardization, comparability, repeatability, traceability) are different and cannot be mixed, even if reprocessed centrally by the ETC as done in this RINGO activity and demonstrated by the results presented in this Deliverable (in particular in the data) and in the analysis included in the Milestone 30.

6 Conclusions

The legacy data reprocessing, starting from the raw measurements, helped to highlight a number of important aspects. Firstly, it is demonstrated that a centralized processing is possible also for other setups and for non-standardized formats of data if the metadata needed for their interpretation and correct evaluation are available. However, it has been also evidenced that the activity could be time consuming and complex in case many different data formats and not well organized metadata are provided. The quality of the results is difficult to be objectively evaluated because there is not a reference target timeseries. The comparison with the PI version is however possible and it evidenced a general agreement with some exceptions in specific years or periods and also function of the setup.

The activity helped also to highlight the main issues that affected the possibility to process the data and the quality of the results. This is also an important information that allowed to 1) confirm that the ICOS strategy followed until now is well designed and all the issues are covered and 2) give information for the importance of the correct collection, organization and storage of these key parameters and characteristics that, if lost make the data impossible to be reprocessed and correctly interpreted.

7 Discussion and future work

One possible strategy in order to have the legacy data included in the products distributed by the ICOS Carbon Portal is to allow Class 1 and 2 stations to prepare and submit a version of the dataset following all the requirements requested to the ICOS Associated station. This could also include a centralized reprocessing of the raw data in order to increase the harmonization and the consistency of the timeseries. This would require however to carefully distinguish the high quality measurements done following the ICOS protocols and the legacy data, that could also include the fluxes measured by the historical system if still running in parallel. On the other side this new product would increase the visibility of ICOS and the long history of measurements that characterizes most of the ICOS ecosystem stations. This possibility will be discussed in the framework of the Monitoring Stations Assembly (MSA) and ICOS RICOM (Research Infrastructure committee) during the first half of 2021 involving also the other components (Ocean and Atmosphere) that could be in a similar conditions with the legacy data.

References

Pastorello, G., Trotta, C., Canfora, E. *et al.* The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. *Sci Data* **7,** 225 (2020). https://doi.org/10.1038/s41597-020-0534-3

Sabbatini, S., Mammarella, I., Arriga, N., Fratini, G., Graf, A., Hörtnagl, L., Ibrom, A., Longdoz, B., Mauder, M., Merbold, L., Metzger, S., Montagnani, L., Pitacco, A., Rebmann, C., Sedlak, P., Sigut, L., Vitale, D., & Papale, D. (2018). Eddy covariance raw data processing for CO_2 and energy fluxes calculation at ICOS ecosystem stations. *International Agrophysics*, *32*(4), 495-515. https://doi.org/10.1515/intag-2017-0043

Sabbatini, S., & Papale, D. (2017). ICOS Ecosystem Instructions for Turbulent Flux Measurements of CO2, Energy and Momentum (Version 20200316). ICOS Ecosystem Thematic Centre. https://doi.org/10.18160/qwv4-639g

Abbreviations

CP	Carbon Portal
CRO	Cropland
EBF	evergreen broadleaf forest
ENF	evergreen needle leaf forest,
EC	Eddy Covariance
ETC	Ecosystem Thematic Centre
FAIR	**F**indable, **A**ccessible, **I**nteroperable, and **R**eusable
L2	Level 2
MSA	Monitoring Stations Assembly
PI	Principal Investigator
PM	Person month
RICOM	Research Infrastructure committee
SN	Serial number
WET	Wetland