ATBD

of HydroGNSS L2OP Surface Inundation and L1OP_CC/CX processors

ESA Contract No. 4000135686/21/NL/AD

SSTL Subcontract No.MS000430

Reference	HydroGNSS-IEEC_L2OPSI-ATBD
Version:	4.7
Prepared by:	E.Cardellach, J. Peng, W.Li
Reviewed by:	M. Unwin
Date:	November 12, 2024
SSTL Ref.:	00649719

History of the document		
Version	**Date**	**Description**
1	Nov. 14, 2021	Original version, including ATBD for all code to be generated at IEEC
1.1	May 5, 2022	ATBD of L2OP-SI processor only. Interfaces of L1b and L2 processor modules provided separately. Further details provided.
1.2	Jun 15, 2022	ATBD updated to answer PDR RIDs on resolution of power spread ratio, maturity of percentage of positive coherence coefficient and quality flags.
2.0	Oct 27, 2022	ATBD updated to account for the improvements in the alpha OP version with respect to the PP version.
2.1	Nov 3, 2022	Clarification about the group of the netCDF output where the output variables are stored.
3.0	Jan 20, 2023	New structure and completion as requested in ESA and SSTL RIDs.
3.1	Feb 8, 2023	Modifications after PDR.
3.2	Mar 8, 2023	More materials on Validation of L2OP-SI and use of different sets of auxiliary data for Random Forest modeling.
4.0	Sep 25, 2023	Document re-structured; introduction of the plans for the use of the 2nd frequency and the 2nd polarization. Track of git tags/hashes and differences with respect to the alpha version.
4.1	Oct 11, 2023	Some corrections, git hash/tag table added.
4.2	Jan 19, 2024	Some corrections, add acronym table.
4.3	Mar 14, 2024	Addressing ESA RIDs MPC-TRR-30, MPC-TRR-31 and MPC-TRR-32.
4.4	May 8, 2024	Update git hash/tag table.
4.5	June 6, 2024	Minor corrections and edits.
4.6	Nov 11, 2024	Clarifications to MPC comments.
4.7	Nov 12, 2024	Definition of the 'inner' core DDM_s in the PSR algorithm. Validation of L2OP-SI based on high sampling rate complex values added. Validation of polarimetric aspects added. New Figure on HSRR added.

EU Export Control	This document and any hardware referred to within has been EU Export Control Rated as:	
Hardware EU Export Rating Number:		Date of rating:
Information EU Export Rating Number:		
I confirm I am an approved Export Rater and this rating is correct as of the date above:		

UK Export Control	This document and any hardware referred to within has been UK Export Control Rated as:	
Hardware UK Export Rating Number:	N/A	Date of rating:
Information UK Export Rating Number:	N/A	27/11/24
I confirm I am an approved SSTL Export Rater and this rating is correct as of the date above:		Martin Unwin

US Export Control	No US Content
ITAR EAR	. N/A
ITAR	
EAR	

1 Applicable Documents **5**
1.1 Acronyms and abbreviations 5
2 Reference Documents **6**
3 Coherence and Surface Inundation Processors **8**
3.1 Sampling rates 10
3.2 Git tags and hashes 10
4 Level 1B processor: coherence detector **12**
4.1 L1B-CX: Coherence detection at low rate from power DDMs 12
4.1.1 Background and scientific justification 12
4.1.2 Algorithm description 12
4.1.3 Validation 16
4.1.4 Roadmap for future improvements 17
4.2 L1B-CC: Coherence detection using complex coherent channel data 17
4.2.1 Background and scientific justification 17
4.2.2 Algorithm description 17
4.2.3 Validation 22
5 L2OP-SI: Inundation processor **26**
5.1 Background and scientific justification 26
5.2 Algorithm description 26
5.2.1 L2-POW: Training of the Random Forest for Incoherent channel 28
5.2.2 L2-CC: Training of the Random Forest for Coherent Channel 30
5.2.3 Interface of L2OP-SI (Beta version) 33
5.3 Validation 34
5.4 Roadmap for future improvements 34
6 Spatial Resolutions **36**
7 Mapping algorithms with Delivered versions and Reference data **37**
Appendix A: INTERNAL TESTS FOR COHERENCE L1B PROCESSOR **39**
Appendix B: VALIDATION OF L2OP-SI **46**
B.1 Validation of the selected approach 46
B1.1 SI from low sampling rate power observables 46
B1.2 SI from high sampling rate complex observables 51
B1.3 Validation of polarimetric aspects 53
B.2 Contribution of auxiliary data 55

1 Applicable Documents

[AD1] HydroGNSS System Requirements Document, ref. 00644076, v1.0, 24/09/2021
[AD2] HydroGNSS Payload Data Ground Segment (PDGS) Algorithm ICD, v13, ref. 00637207V13, 07/12/2023
[AD3] ESA Scout-2 HydroGNSS SRD Science Team Annex, ref. 00635466v3 HydroGNSS SRD ScienceTeam Annex, v3, 01/10/2021
[AD4] Scout-2 HydroGNSS VCD for IEEC, ref.00644970, v1.0, 23/01/2023

1.1 Acronyms and abbreviations

The following abbreviations are used within this document:

ATBD	Algorithm Theoretical Baseline Document
BF1A/B	BuFeng 1A/B
CCI	Climate Change Initiative
CyGNSS	Cyclone Global Navigation Satellite System
DDM	Delay Doppler Map
E2ES	End To End Simulator
ESA	European Space Agency
FSScat	Federated Satellite Systems 3Cat
FY3E	FengYun-3E
GNSS-R	Global Navigation Satellite System Reflectometry
GSW	Global Surface Water
IEEC	Institut d'Estudis Espacials de Catalunya
L1B-CC	L1B Processor for Coherent Channel
L1B-CX	L1B Processor for Incoherent Channel
L1OP	Level 1 Operational Processor
L2OP	Level 2 Operational Processor
L2OP-SI	Level 2 Operational Processor for Surface Inundation

LHCP	Left Hand Circular Polarization
PDGS	Payload Data Ground Segment
PSR	Power Spread Ratio
RHCP	Right Hand Circular Polarization
SNR	Signal to Noise Ratio
SP	Specular Point
TDS-1	TechDemoSat-1

2 Reference Documents

[AK21] Mohammad M. Al-Khaldi, Joel T. Johnson , Scott Gleason, Eric Loria, Andrew J. O'Brien, and Yuchan Yi, An Algorithm for Detecting Coherence in Cyclone Global Navigation Satellite System Mission Level-1 Delay-Doppler Maps, IEEE TGRS, VOL. 59, NO. 5, MAY 2021 doi:10.1109/TGRS.2020.3009784

[AK21b] Mohammad M. Al-Khaldi, Rashmi Shah, Clara C. Chew, Joel T. Johnson, and Scott Gleason. Mapping the Dynamics of the South Asian Monsoon Using CYGNSSs Level-1 Signal Coherency. IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, 14:1111–1119, 2021.

[A18] Amatulli, G., Domisch, S., Tuanmu, M.-N., Parmentier, B., Ranipeta, A., Malczyk, J., and Jetz, W. (2018) A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. Scientific Data volume 5, Article number: 180040. DOI: doi:10.1038/sdata.2018.40.

[B19] Ahmed M. Balakhder, Mohammad M. Al-Khaldi, and Joel T. Johnson. On the Coherency of Ocean and Land Surface Specular Scattering for GNSS-R and Signals of Opportunity Systems. IEEE Transactions on Geoscience and Remote Sensing, 57(12):10426–10436, December 2019.

[C20] Adriano Camps. Spatial Resolution in GNSS-R Under Coherent Scattering. IEEE Geoscience and Remote Sensing Letters, 17(1):32–36, January 2020.

[D21] Zhounan Dong and Shuanggen Jin. Evaluation of the Land GNSS-Reflected DDM Coherence on Soil Moisture Estimation from CYGNSS Data. *REMOTE SENSING*, 13(4), February 2021.

[E17] ESA, Land Cover CCI Product User Guide Version 2. Tech. Rep. (2017). Available at: maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf

[L18] Weiqiang Li , Antonio Rius , Fran Fabra , Estel Cardellach , Serni Ribó , and Manuel Martín-Neira, Revisiting the GNSS-R Waveform Statistics and Its Impact on Altimetric Retrievals, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING 2018, doi:10.1109/TGRS.2017.2785343

[L21] Weiqiang LI, Estel CARDELLACH, Serni RIBÓ, Antonio RIUS, Bo ZHOU, First spaceborne demonstration of BeiDou-3 signals for GNSS reflectometry from CYGNSS constellation, Chinese Journal of Aeronautics, (2021), 34(9): 1–10, doi:10.1016/j.cja.2020.11.016.

[L22b] Li, W.; Cardellach, E.; Ribó, S.; Oliveras, S.; Rius, A. Exploration of Multi-Mission Spaceborne GNSS-R Raw IF Data Sets: Processing, Data Products and Potential Applications. Remote Sens. 2022, 14, 1344. doi:/10.3390/rs14061344.

[L22] Breiman, Leo. "Random forests." Machine learning 45 (2001): 5-32.

[L20] Eric Loria, Andrew O'Brien, Valery Zavorotny, Brandi Downs, and Cinzia Zuffada. Analysis of scattering characteristics from inland bodies of water observed by CYGNSS. *Remote Sensing of Environment*, 245:111825, August 2020.

[M20] Joan Francesc Munoz-Martin, Raul Onrubia, Daniel Pascual, Hyuk Park, Adriano Camps, Christoph Rudiger, Jeffrey Walker, and Alessandra Monerris. Untangling the Incoherent and Coherent Scattering Components in GNSS-R and Novel Applications. *Remote Sensing*, 12(7):1208, April 2020.

[P16] Pekel, J., Cottam, A., Gorelick, N. and Belward, A., Global Surface Water - Data Access, Nature, 2016, doi:10.1038/nature20584, JRC109054.

[P24] Peng, J., W. Li, E. Cardellach, G. Marigold, M.P. Clarizia, *Signal coherence and water detection algorithms for the ESA HydroGNSS mission*, IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1-18, 2024, Art no. 5801218, doi: 10.1109/TGRS.2024.3410876

[P25] Peng, J., E. Cardellach, W. Li, S. Ribó, A. Rius, *Towards the exploitation of HydroGNSS coherent channel: High sampling rate products*, in preparation.

[P25b] Peng, J., E. Cardellach, W. Li, S. Ribó, A. Rius, *Impact of right-hand polarized signals in GNSS-R water detection algorithms*, under review at IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.

3 Coherence and Surface Inundation Processors

This document describes the algorithms developed to detect coherence and surface water in HydroGNSS measurements. These algorithms can be classified according to the level (level-1b vs. level-2), or according to the type of measurement used as input (low rate power measurements vs. high rate complex signals). Furthermore, the algorithms can be applied to different signals, that is, signals from different frequency bands and different polarizations.

The overall strategy, when all signals will be available, is shown in Figure 3a. The initial input consists of low rate power DDM and high rate coherent channel complex signals from all frequency-polarization combinations (L1-LH, L1-RH, L5/E5-LH, L5/E5-RH). The final output consists of two independent solutions of along-track water percentages and flags: one unique track at low sampling rate and another unique track at higher sampling rate. A discussion about the terminology that will be used afterwards regarding the sampling rates is provided later in this section (see Section 3.1).

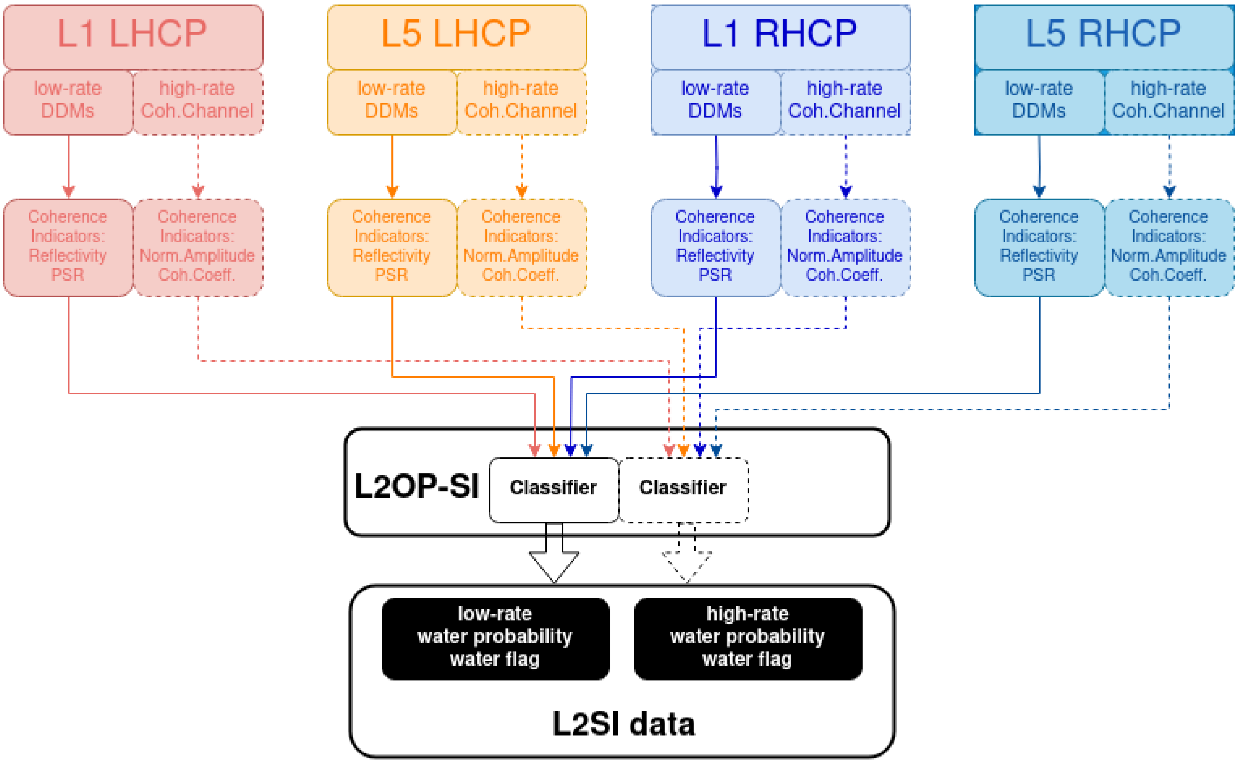


Figure 3a: *Overall approach to obtain one along-track solution of the water inundation at low rate, and another one at higher rate. These final water probabilities are obtained from two types of data, the processing chain of each type being independent from the other: the power DDM processing chain is indicated in solid lines whereas the coherent channel complex signals processing chain is shown in dashed lines. The L2OP-SI processor includes both power and coherent channel algorithms to generate a single L2-SI file (low-rate and high-rate water information in different groups of the file). Figure adapted from [P24].*

The current implementation only uses L1-LHCP signals. The same algorithms to detect coherence (L1B-CX and L1B-CC) will be applied to the other frequency and polarization signals when available. The water classifiers (one for low rate power measurements and another for high rate complex signals) will be re-trained to combine all frequencies and polarizations in a single along-track water solution (one at each sampling rate). While waiting for the dual-frequency dual-polarization data, the current implementation with L1-LHCP is shown in the block diagram in Figure 3b below.

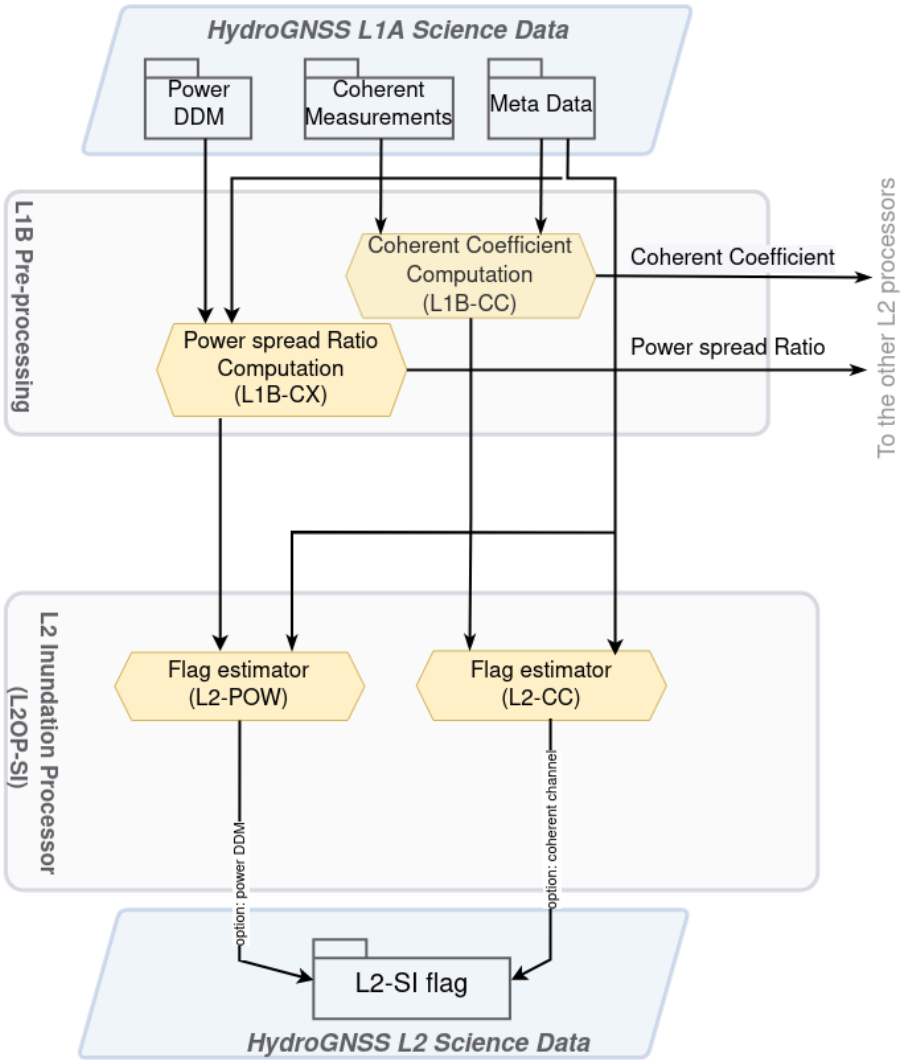


Figure 3b: *Block diagram of the current implementation of the L2OP SI processing chain (including L1B coherence detection algorithms). This only uses L1-LHCP signals. See Figure 3a for the approach with the complete set of HydroGNSS signals.*

The overall task is split in different processors, each one works as a stand-alone piece of code and are provided separately. They are indicated in orange hexagons in Figure 3b. At the moment (beta version), the four of them are ready and have been delivered: two level-1b processors to detect coherence, one from DDM power information (L1B-CX) and the other from coherent channel complex information (L1B-CC), and one level-2 processor that identify the surface water or inundation on the power-processor chain (L2-POW) and the on the complex signals of the coherent channel (L2-CC). The part of the coherent channel processing chain (L2-CC) has been trained with an insufficient amount of data but it has already been delivered. The output of the level-1 processors is included in the L1B data, while the output of the level-2 processor is included in the L2 products.

The algorithms and their validation are described in the peer-reviewed article *Peng et al., 2023* [P23] – submitted to IEEE TGRS, resolution pending.

The level-1b processors are explained in Section 4 and the level-2 algorithms in Section 5.

3.1 Sampling rates

These processors deal with 3 different sampling rates, as described in Table 3.1.

***Table 3.1**: Description of the three different sampling rates involved in these algorithms.*

Very-high sampling rate:	Sampling given by the coherent channel of the receiver, with coherent integration time t_c of one to a few milliseconds. **Baseline t_c= 4 ms → sampling 250 Hz.**
High sampling rate:	Sampling at the end of the L1B complex coherent channel processor. It corresponds to a coherent integration time of ten/s of milliseconds, the result of extending the coherent integration by M_c samples originally integrated during t_c batches, so $T_c = t_c \times M_c$. **Baseline: T_c= 4 ms x 10 samples = 40 ms → sampling 25 Hz.**
Low sampling rate:	Sampling provided by the power DDM L1 observables, after coherent and incoherent integration (total integration T_{in}). **Baseline: T_{in} = 1 or 0.5 s → sampling 1 or 2 Hz.**

3.2 Git tags and hashes

***Table 3.2:** Git tags and hashes of different versions of the processors.*

Processor ID	Version:	Delivery date:	Git tag or hash
L1B-CX-CC	alpha	01/23/2023	HydroGNSS-IEEC_L1B_CX-CCC
L2-SI	alpha	01/23/2023	HydroGNSS-IEEC_L2OPSI-PSR v2.0
L1-CX-CC	beta	05/10/2023	4c4bab7a
L2-SI	beta	05/10/2023	d3a6f59b

L1-CX-CC	beta	09/10/2023	b872ce6e
L2-SI	beta	09/10/2023	dcf3ac45
L1-CX-CC	beta	12/22/2023	9576a2aa
L2-SI	beta	12/22/2023	96658b12
L1-CX-CC	beta	04/16/2024	e600c7ab
L2-SI	beta	04/18/2024	1a401c94
L1-CX-CC	beta	05/24/2024	43de4869
L2-SI	beta	05/24/2024	f93d77b7
L1-CX-CC	beta	11/12/2024	5de37ce4
L2-SI	beta	11/12/2024	c15e64f7

4 Level 1B processor: coherence detector

There are two Level 1B coherence detectors: one that works at low sampling rate (1-2 Hz) on power DDM observables (processor called L1B-CX) and another processor that works at high sampling rate (250-50 Hz) on complex coherent channel observables (L1B-CC). These algorithms are described in Sections 4.1 and 4.2 respectively. At the moment, they have been validated with L1-LHCP data only, the processors will be validated using dual frequency and dual polarization data from payload commissioning collection phase after HydroGNSS launch.

4.1 L1B-CX: Coherence detection at low rate from power DDMs

4.1.1 Background and scientific justification

The detection and quantification of the coherent component in GNSS-R scatterometry is a research line that has attracted some attention recently. At the moment this contract kicked off, a few works had been published [B19, L20, C20, M20, AK21, AK21b, D21]. Due to the requirement on rapid implementation of HydroGNSS, the selection of a published algorithm for coherence detection could facilitate the reduction of the development time, thus respecting the delivery plans. Among the published ones, IEEC team selected [AK21] for its relative simplicity and encouraging results.

4.1.2 Algorithm description

INPUT

- Level L1b power DDMs
- Metadata

STEPS

The first step is to search for the actual peak power, and identify the delay-Doppler cell where it lays. This is done with python built-in functions (`numpy.argmax`). Note that the search is conducted across all DDM cells, irrespectively of the distance to the estimated specular point cell (a flag is provided to warn about peaks identified too far away from the nominal specular point).

Once the peak is located, a subset of delay-Doppler cells around it are selected. We will call this subset of cells the 'DDM zone of interest' (DDM_S). DDM_S is the piece of DDM that will be used to compute the coherence indicator. The delay-Doppler cells around the peak used for these computations cover a range of Doppler frequencies +- 2500 Hz from the peak (so 5 kHz in total) and a range of delays +-

1995.625 nanoseconds (approximately +- 600 m from the peak). These values are stored in the variables DopplerRangeHz and DelayRangeNanoSec of the L1B-CX algorithm and could be changed, if needed.

These values determine the number of delay lags and Doppler slices around the peak that are used for the computations. These are:

$$NHalfDoppler \; = int(\frac{DopplerRangeHz}{DopplerResolution})$$
$$NHalfDelay \; = \; int(\frac{DelayRangeNanoSec}{DelayResolution})$$

where DopplerResolution and DelayResolution are the Doppler and delay spacing in the original data, in Hz and nanoseconds respectively.

Therefore, NHalfDoppler and NHalfDelay are the number of Doppler and delay cells between the peak power and the edge of the area of interest, DDM_S.

With the current settings and values provided above, and assuming 226.44 nanoseconds DelayResolution and 500 Hz DopplerResolution, NHalfDoppler = 5, NHalfDelay = 8, so the size of area of interest, DDM_S is 11 Doppler x 17 delay cells (with CYGNSS data, it corresponds to 7 Doppler x 11 delay cells).

At this stage, there are two things that might hinder the interpretation of the outcome of the computations:

- The area of interest, DDM_S, is partially out of the actual measured DDM. When this happens, the computations of the coherence indicators could not be compared to those done over a complete DDM_S. Then, the coherence indicator computations are not performed, they are set to zero, and a flag called **'DDMPeakQualityFlags'**, also set to one, indicates that the coherence indicator should not be used. This is illustrated in Figure 4.1.2, cases B and C. Otherwise, the **'DDMPeakQualityFlags'** is set to zero, meaning that this problem does not apply.
- It might also happen that the area of interest, DDM_S, is fully inside the measured DDM, but the distance of the peak to the nominal specular point is large enough so the DDM_S does not cover the area of the nominal specular point (so, on the land surface, the coherence computations will be unrelated to the coordinates where the specular point is expected). To give an indication on whether the coherence indicators and flag might not inform about the nominal specular point location, a flag called **'DDMPeakOffsetDistance'** contains the distance, in delay-Doppler cells, between the cell corresponding to the nominal specular point and the peak power one. As illustrated in Figure 4.1.2, case A would correspond to this case, with **'DDMPeakOffsetDistance'** = 14.

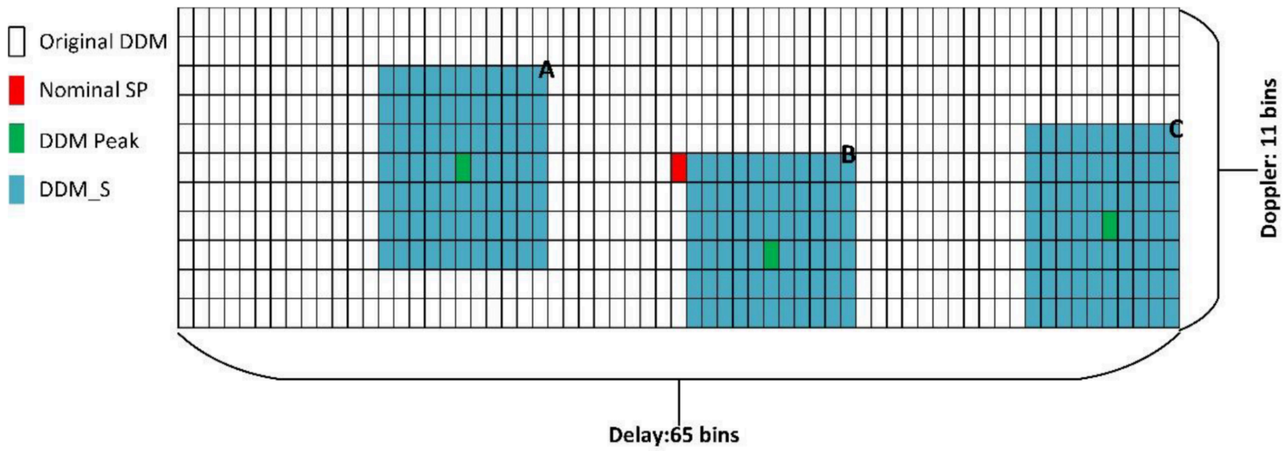


Figure 4.1.2: *Sketch of the delay-Doppler cells in a DDM formatted as CYGNSS (white boxes), with the nominal specular point located at the red cell, and three different examples of* DDM_S *(blue shade of 11x7 cells) where the coherence detection algorithm should be disregarded or used with caution. Case A: the peak power, located in the green center of the blue zone (*DDM_S *zone), is fully inside the DDM measurements grid, but the information contained in this area does not include information around the nominal (and probably actual) specular point. This is informed through the 'DDMPeakOffsetDistance' flag. Cases B and C: the irrespective peak power cells are too close to the edge of the DDM that the area of interest (blue cells) are not fully measured. In these cases, the computation of the coherence indicators could not be compared to the regular ones and the computations do not take place, they are set to one together with the flag 'DDMPeakQualityFlags', also set to one.*

The coherence indicator is the Power Spread Ratio (PSR), from [AK21]. The **'PowerSpreadRatio'** is the ratio between the sum of powers in the 'inner' core of the DDM_s and the sum of powers above the noise level in the DDM_s. The noise level is here defined as in [AK21, Fig.2]. The 'inner' core is currently defined as ±250 ns delay x ±1000 Hz Doppler centered at the peak. These values are called the DelayRangeNanoIn and the DopplerRangeHzIn. DelayRangeNanoIn and the DopplerRangeHzIn are set in the 'init' function of the class L1B_CX, and mapped into delay-Doppler pixels as:

$$NHalfDopplerIn = round\left(\frac{DopplerRangeHzIn}{DopplerResolution}\right)$$

$$NHalfDelayIn = round\left(\frac{DelayRangeNanoIn}{DelayResolution}\right)$$

Then, the power spread ratio is:

$$PSR = \frac{C_{in}}{C_{out}}$$

$$C_{in} = \sum_{i=-NHalfDelayIn}^{NHalfDelayIn} \sum_{j=-NHalfDopplerIn}^{NHalfDopplerIn} DDM(\tau_P + i, f_P + j)$$

$$C_{out} = \sum_{i=-NHalfDelay}^{NHalfDelay} \sum_{j=-NHalfDoppler}^{NHalfDoppler} DDM(\tau_P + i, f_P + j) - C_{in}$$

In the equations above,

- The central pixel is the one of the peak power (τ_P, f_P), not the nominal specular point;
- C_{out} only accounts for cells within the DDM_S zone (blue zones in Figure 4.2.1) but the cells below the noise level are excluded from the C_{out} summation (following [AK21, Fig.2]).

Most scattering events can present both coherent and diffuse components to the total scattered field. A threshold has been defined (PSRThreshold) to determine whether the event presents predominance of coherent scattering. That is, **events with power spread ratio PSR >= PSRThreshold are labeled as 'coherent'**, as described in Table 4.1.2a. This 'coherence flag' is also an output of the algorithm, called '**Coherency**' in the L1B data. The nominal value of PSRThreshold is 2.

Table 4.1.2a: *Thresholds applied to flag a power DDM observation as 'coherent'.*

Coherence Flag (L1B-CX)	
Incoherent (coherent flag = 0)	**Coherent (coherent flag = 1):**
Power spread ratio PSR < PSRThreshold	Power spread ratio PSR >= PSRThreshold

The Power Spread Ratio PR and the coherence flag are both provided in the level 1b (L1B) metadata, under 'incoherent' groups, together with two quality flags.

The quality flags provided with the PSR value relate to the location of the delay-Doppler cells used to compute it, as already detailed above. These flags do not inform about the general quality of the observables, because these are already provided by other L1 algorithms (e.g., low SNR). A summary is given below:

- **'DDMPeakQualityFlags'**: when some of the cells of the N_τ x N_f box required for PSR computations lay out of the actual DDM, then a warning flag is provided, as it might alter the results.
- **'DDMPeakOffsetDistance'**: Similarly, a DDM that presents its peak power away from the nominal specular point (0,0) delay-Doppler cell might correspond to coherent scattering off topographic features that are not located at the specular point. Therefore, despite the scattering

captured by the DDM might be coherent, this might not be located at the coordinates given in the metadata.

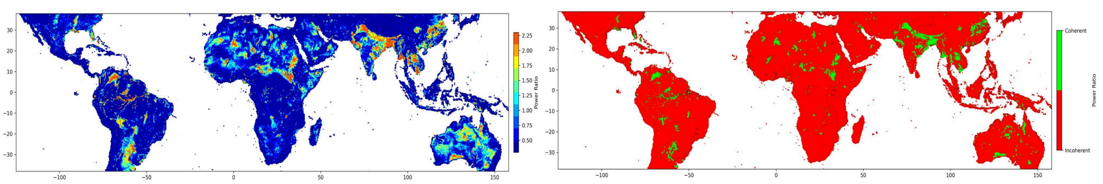
Table 4.1.2b: *Interface of the L1B-CX processor algorithm. The user only needs to call it after changing the input path in the configuration file, but the inner algorithm is called with particular inputs and outputs as described below.*

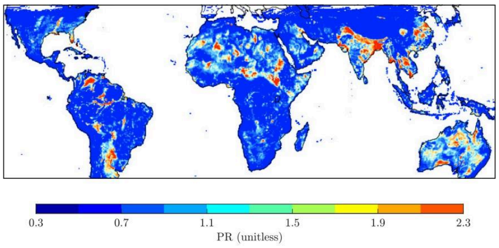
Interface of L1b pre-processor algorithm using power DDM	
SYNTAX (full directory):	L1B_CX_DR.exe
INPUT	**OUTPUT (incoherent groups)**
-Level L1b power DDMs -Metadata	-'**PowerSpreadRatio**': Power Spread Ratio (PSR) @ T_{incoh} rate [32-bit floating point] -'**Coherency**': Coherence Flag @ T_{incoh} rate [16-bit unsigned integer] -'**DDMPeakQualityFlags**': Quality Flag of the PSR and Coherence Flag products, one when part of the DDM box to be used for the PSR computation lays out of the actual DDM, @ T_{incoh} rate [16-bit unsigned integer] -'**DDMPeakOffsetDistance**': quality indicator of the PSR and Coherence Flag that warns of actual distance (in delay-Doppler cells) between the actual DDM peak and the nominal specular point delay-Doppler cell location. This can be an indicator of coherent reflection being produced in topographic features away from the specular point. [16-bit unsigned integer]
SAMPLING RATE:	Low sampling rate (baseline 1 or 2 Hz).
TYPE:	Along track product, no gridding, no interpolation.

4.1.3 Validation

A qualitative comparison was conducted between the Figures in [AK21] and the PSR computed by the L1B-CX algorithm applied to the same CyGNSS data sets, as shown in Figure 4.1.3.

See also the validation of the L2 products produced with the PSR coherence indicator, in Appendix B.





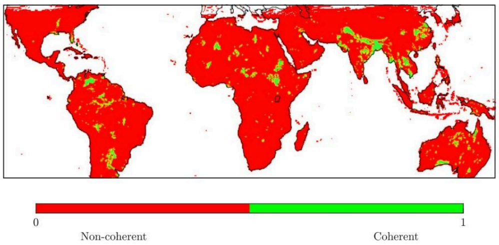


Figure 4.1.3: *(Left) PSR. (Right) Coherence Flag. (Top) Results from L1B-CX algorithm. (Bottom) Figures from [AK21]. All figures have been generated with CyGNSS Level 1 Science Data Record Version 3.1 for the same period (DOY: 240-251 in 2019).*

4.1.4 Roadmap for future improvements

HydroGNSS will collect complex data through its coherent channel, simultaneously to the traditional power observables used for the PSR algorithm. Coherence detection will benefit from the use of both algorithms simultaneously and at different sampling rates.

HydroGNSS will collect data at different combinations of polarization and frequency band. The L1B-CX algorithm will be applied to all these combinations. Modifications will be done if required – e.g., threshold of PSR>2 used in L1-LHCP might need to be different in other frequency-polarization signals.

4.2 L1B-CC: Coherence detection using complex coherent channel data

4.2.1 Background and scientific justification

HydroGNSS will be the first near-nadir GNSS-R mission provided with the 'coherent channel'. Previous GNSS-R missions (e.g., TDS-1, CyGNSS, FY3E, BF1A/B, FSScat), they all collected incoherently integrated (power) measurements, losing the complex (phase) information. The exceptions were the seldom and short raw intermediate frequency (IF) sampled signal acquisitions, which permitted, once on the ground, to process in-lab and generate any low level products. For this reason, at the time of kicking off this contract, there was no literature about the use of the 'coherent channel' for coherence detection, but our work using raw IF sampled signals [L21, L22b]. This is the approach implemented in the L1-CC algorithm, described below.

4.2.2 Algorithm description

INPUT

- Level L1b coherent channel
- Metadata

STEPS

In this processing chain, the input data corresponds to a very-high sampling rate (e.g., 250 Hz or 4 ms integration), and the output data corresponds to a high sampling rate (e.g., 25 Hz or 40 ms coherent integration), as described in Table 3.

The coherence indicator computed with very high sampling rate complex I/Q data is the coherent coefficient described in [L21, L22b], called **'CoherenceCoefficient'** in the output L1B data file. First, the coherent coefficient (r_{coh}) of two samples of the coherent channel is computed, this uses the i-element and the element that is delayed by N_c samples (i+N_c, default N_c=1):

$$r^i_{coh} = \frac{z^i_r\left(\tau_p, f_p\right) z_r^{(i+N_c)*}\left(\tau_p, f_p\right)}{\left|z^i_r\left(\tau_p, f_p\right)\right| \left|z_r^{(i+N_c)}\left(\tau_p, f_p\right)\right|}$$

Where the index 'i' is associated at the 'very-high sampling rate' time series (original output of the coherent channel, see Table 3).

Secondly, M_c subsequent r_{coh} are coherently averaged, and the final amplitude is the coherence indicator for the coherent channel R_{coh}^k . The number of samples M_c is computed from the TargetCohResolution, that is, the total final coherent temporal integration time of the L1B complex products to be achieved. The nominal value for TargetCohResolution is 40 milliseconds, which leads to M_c=10 samples at t_c=4 ms integration each each sample. The t_c values cannot be set in this algorithm, but it is a characteristics of the data (receiver settings).

Therefore, the coherence indicator obtained with the coherent channel has a sampling time of $T_c=M_c*t_c$, what we called 'high sampling rate' in Table 3. This coherence indicator at high sampling rate, R_{coh}, is called '**CoherenceCoefficient**' in the L1B data.

$$R^k_{coh} = \frac{\left|\sum_{i=kM_c}^{(k+1)M_c} r^i_{coh}\right|}{M_c}$$

Note that R_{coh}^k is a value between 0 and 1, as it was produced by the summation of a number M_c of unitary phasors r_{coh} divided by M_c. This coherence coefficient will only have value 1 when all r_{coh} are perfectly aligned in the complex space, meaning that the phase jumps between any i-sample and the i+N_c sample all have the same phase leap (e.g., as in a perfect tone).

Because the sampling rate is different, the L1B output file must also provide the time-tag of this new time series, called '**Time_CoherentInt**' in the L1B files. Similarly, the algorithm also provides the corresponding specular point locations and incidence angles, all interpolated (third order spline interpolations) at '**Time_CoherentInt**' and given in the L1B files as '**Longitudes_CoherentInt**', '**Latitudes_CoherentInt**' and '**Incidence_CoherentInt**'.

The threshold for 'CoherenceCoefficient' is the parameter CoherenceFlagThreshold, that is, the value of the Coherent Coefficient above which signals are assumed with predominance of coherent scattering and below which signals are considered incoherent. This value is nominally at 0.6.

That is, events with Coherence Coefficient >= CoherenceFlagThreshold are labeled as 'coherent', as described in Table 4.2.2a. This flag, called **'WF_CoherenceFlag'** is also an output of the algorithm. This value is selected based on the R_{coh} values found in CyGNSS data (from raw IF acquisitions) as co-located with the Global Surface Water data sets, as seen in Figure 4.2.2a.

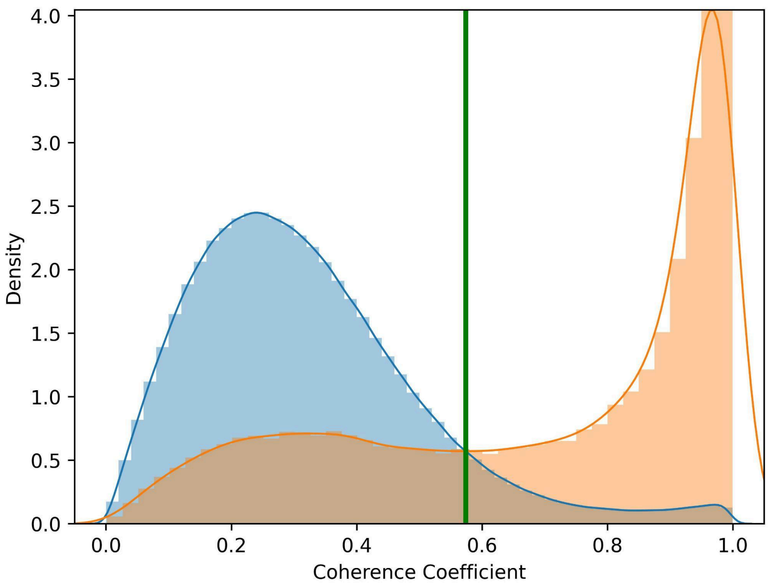


Figure 4.2.2a: *Distribution of 'CoherenceCoefficient' R_{coh} computed for CyGNSS data obtained in raw IF acquisition mode and processed on ground, at IEEC software receivers. The blue corresponds to samples over dry land and orange for samples over inundated surface, as informed by the reference data set Global Surface Water.*

Table 4.2.2a: *Thresholds applied to flag the complex coherent channel measurements as 'coherent'.*

Coherence Flag (L1B-CC)	
Incoherent (coherent flag = 0)	**Coherent (coherent flag = 1):**
Coherence Coefficient < CoherenceFlagThreshold	Coherence Coefficient >= CoherenceFlagThreshold

The quality flag of the R_{coh} product identifies discontinuities (first and last samples of the continuous time series, where the M_c windows might not be complete), as this alters the results of the algorithm, which depend on M_c neighboring samples.

Additionally, a low rate indicator is defined at the percentage of Abs{R_{coh}} with values above 0.6, in time batches that correspond to the incoherence integration in the power DDMs. That is, percentage of coherent high sampling rate samples in one incoherent-time batch:

$$Pcc = 100 * number\ (Abs\{R_{coh}\} > 0.6)_{[Tincoh]} \ /\ number\ coh\ channel\ samples\ in\ Tincoh$$

Studies with synthetic data show that P_{cc} values above 60% correspond to signals with significant coherent components, as shown in Figure 4.2.2b. Appendix A compiles some of the studies conducted with synthetic data.

Another coherence indicator is the Normalized Amplitude NA_{coh}^k , which is obtained by normalizing the amplitude of mean complex I/Q data through the '**CoherenceCoefficient**'.

$$A_{coh}^k = \left| \sum_{i=kM_c}^{(k+1)M_c} z_r^i(\tau_p, f_p) \right|$$

$$NA_{coh}^k = \frac{A_{coh}^k - min(A_{coh}^k)}{\Delta A_{coh}^k} \Delta R_{coh}^k + min(R_{coh}^k)$$

Note that ΔA_{coh}^k and ΔR_{coh}^k are the difference between the maximum in the track and the minimum in the track of amplitude and coherence coefficient, respectively.

The High Sampling Rate Reflectivity $HSRR_{coh}^k$ is obtained by an exponential function fitting of the mean normalized amplitude and the reflectivity from the incoherent channel.
The steps to compute HSSR are as follow:

1. Average the Normalized Amplitude up to 1 second sampling (NA_{1Hz}).
2. Fit an exponential function between the 1-second averaged NA and the 1-second L1B Reflectivity given in dB (from low rate DDMs data), so, A, B and C are the best fit for $\Gamma_{1Hz,dB} = A\ e^{B \cdot NA_{1Hz}} + C$. This is done for each individual track.
3. Apply the exponential function to the high sampling rate normalized amplitudes, also given in dB:

$$HSRR_{coh}^k = A\ e^{B \cdot NA_{coh}^k} + C$$

Please note that HSRR is neither used for the coherence nor the water detection algorithms described in this document. This is an experimental intermediate product (L1b) provided for the users to play. It should be equivalent (and 'calibrated' as) the low sampling rate reflectivity, but at a higher rate. At this stage, it has not been fully demonstrated whether the high rate signatures are indeed geophysical, but it is provided for research purposes [P25].

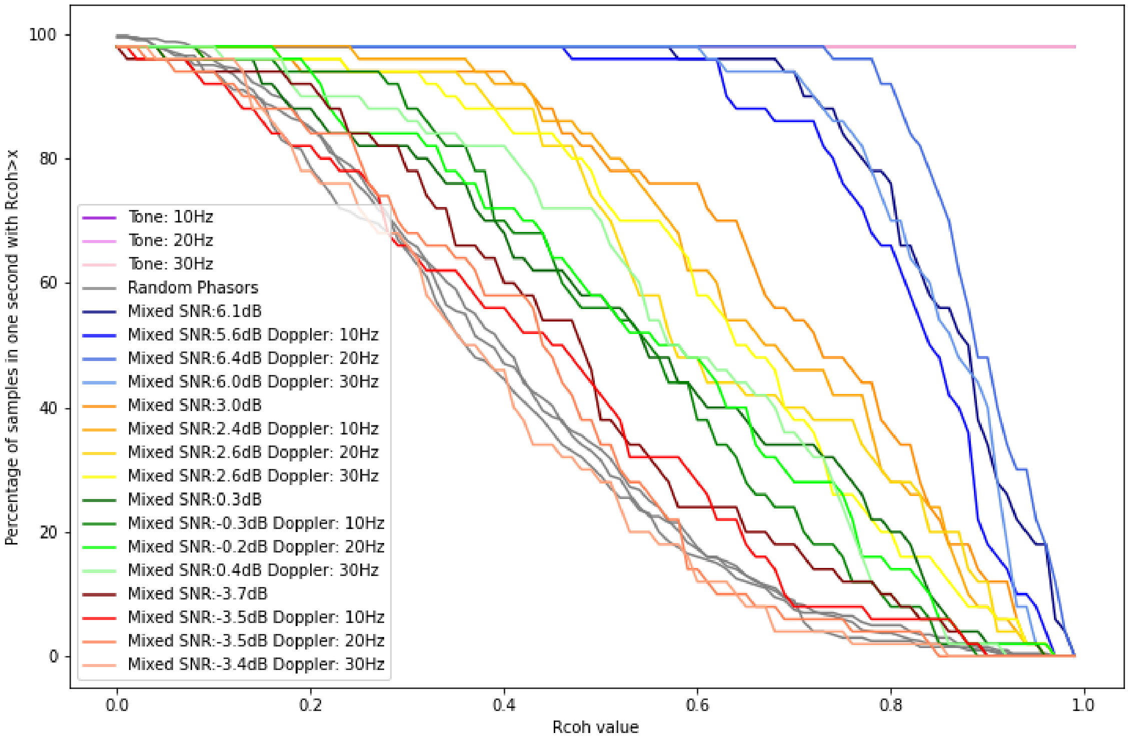


***Figure 4.2.2b**: Results of the Pcc algorithm applied on a series of synthetic data. Each color hue relates to the SNR values, while variations in the same hue relate to residual dynamics (residual Doppler frequencies that make the phasor spin). In grey, different realizations of time series of pure random phasors, and in magenta three purely coherent tones, but with different residual dynamics – their results overlap.*

The integrated Coherent Coefficient R_{coh} and the Percentage P_{cc} are included in the level-1b data (L1B).

***Table 4.2.2b**: Interface of the L1B-CC processor algorithm. The user only needs to call it with a path to the L1 files, but the inner algorithm is called with particular inputs and outputs as described below.*

Interface of L1b pre-processor using coherent channel	
INPUT	**OUTPUT (coherent groups)**
-Level L1b coherent channel -Metadata	-'**CoherenceCoefficient**': Amplitude of the integrated coherence coefficient R_{coh} after $M_c T_{coh}$ integration (e.g., M_c = 10) [32-bit

	floating point] -'**WF_CoherenceFlag**': Coherence Flag of the R_{coh} product [16-bit unsigned integer] -'**PercentageCoherenceCoefficient**': Percentage of coherent samples (WF_CoherenceFlag=1) in $T_{incoheren}$, P_{cc} (typically 1 second) @ T_{incoh} rate [32-bit floating point] -'**Time_CoherentInt**': Time of the integrated coherence coefficient R_{coh} [64-bit floating point] -'**Longitudes_CoherentInt**': Longitude of the integrated coherence coefficient R_{coh} [32-bit floating point] -'**Latitudes_CoherentInt**': Latitude of the integrated coherence coefficient R_{coh} [32-bit floating point] -'**Incidence_CoherentInt**': Incidence of the integrated coherence coefficient R_{coh} [32-bit floating point] -'**NormalizedAmplitude**': Normalized amplitude of the complex I/Q data [32-bit floating point] -'**HighSamplingRateReflectivity**': High sampling rate reflectivity of the complex I/Q data [32-bit floating point]
SAMPLING RATE:	R_{coh} : high sampling rate, 25 Hz baseline, it can be deduced from 'Time_CoherentInt' [32-bit floating point] P_{cc} : low sampling rate, 1 Hz baseline, same as the incoherent measurements in L1A data [32-bit floating point]
TYPE:	Along track product, no gridding. Decimation in time.

4.2.3 Validation

The limited amount of data hinders the validation of the coherent channel products. A qualitative comparison was conducted between the Figures in [L22b] and the CoherenceCoefficient computed by the L1B-CC algorithm applied to the same CyGNSS data sets, as shown in Figure 4.2.3a.

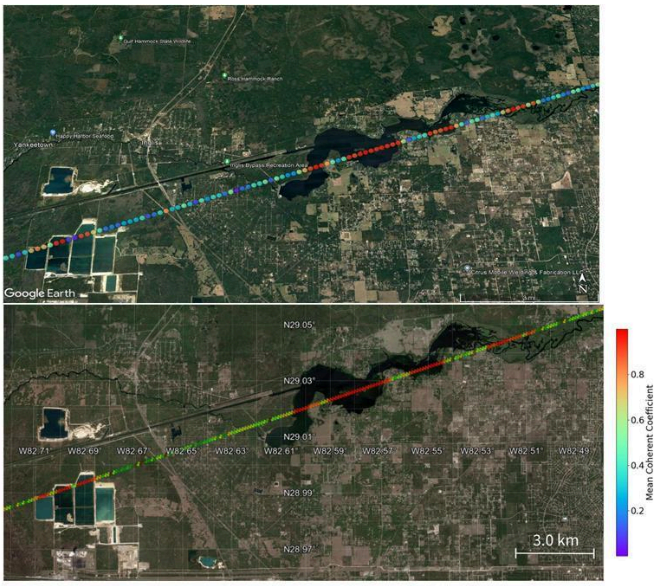


Figure 4.2.3a: *(Top) Results from L1B-CC algorithm. (Bottom) Figures from [*L22b*]. The raw IF data was collected by CYGNSS FM02 on 13 September 2018 over Florida, US.*

A qualitative comparison was conducted between the Normalized Amplitude and the Coherence Coefficient computed by the L1B-CC algorithm applied to the same CyGNSS data sets, as shown in Figure 4.2.3b.

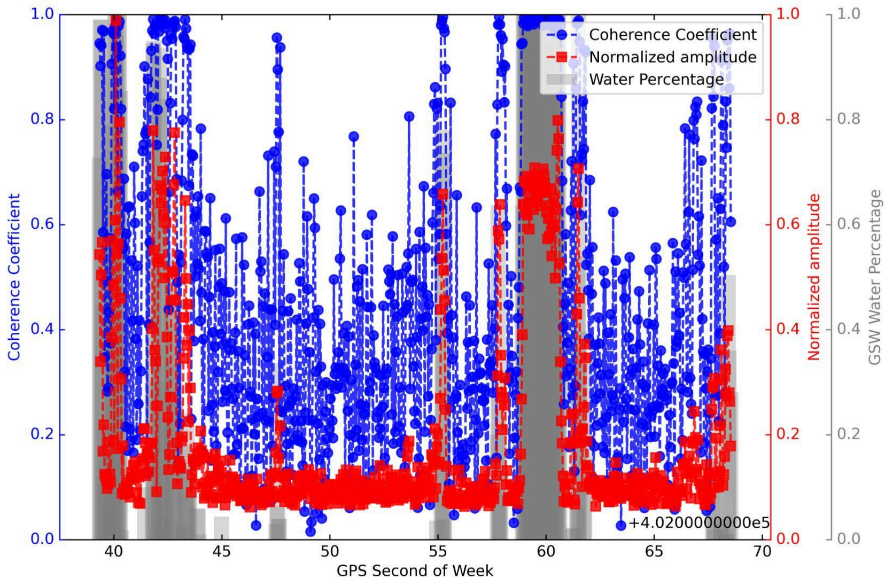


Figure 4.2.3b: *The comparison between coherence coefficient and normalized amplitude. The raw IF data was collected by CYGNSS FM02 on 13 September 2018 over Florida, US.*

Figure 4.2.3c illustrates a comparison between the Reflectivity obtained from the incoherent channel and the High Sampling Rate Reflectivity computed using the L1B-CC algorithm on identical CyGNSS datasets, while Figure 4.2.3d shows it for two CYGNSS tracks nearly overlapping and acquired within

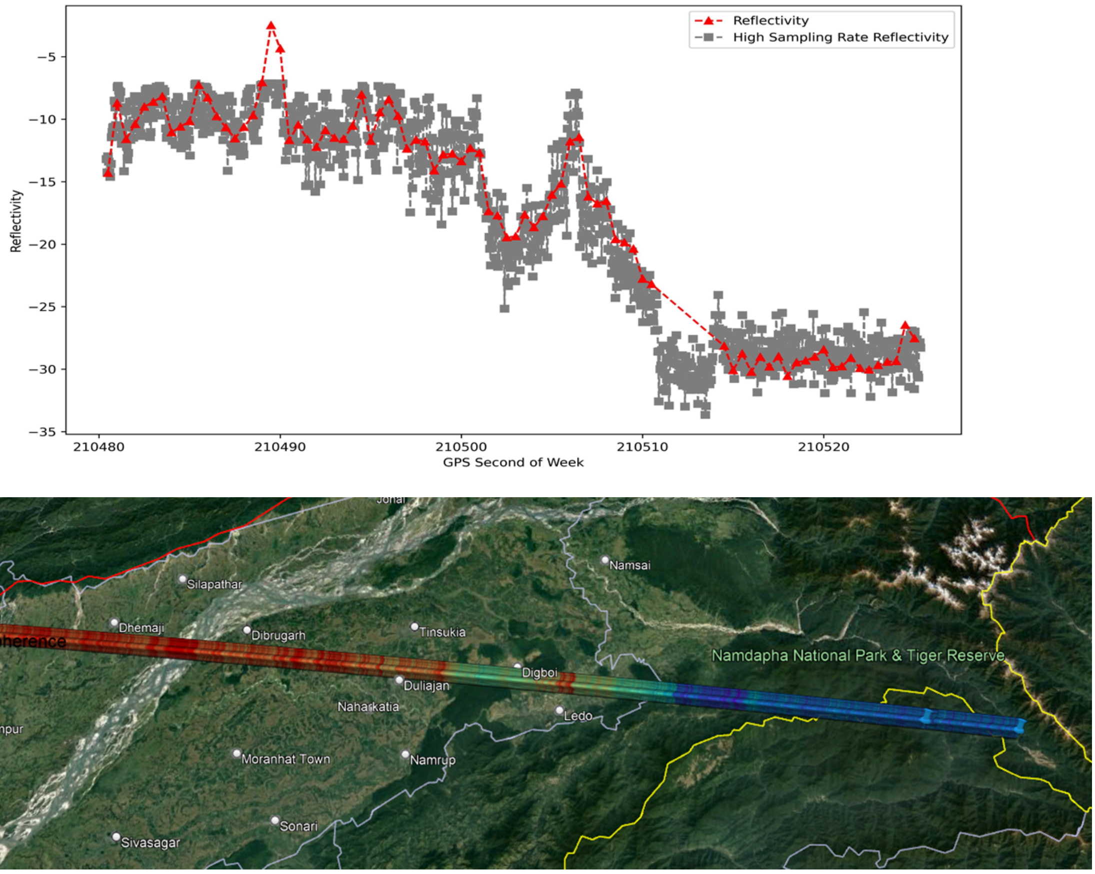


Figure 4.2.3c: *(Top) HSRR results from L1B-CC algorithm and incoherent channel. (Bottom) The result displayed on Google earth. The raw IF data was collected by CYGNSS FM06 on 20 August 2019 over Assam, India.*

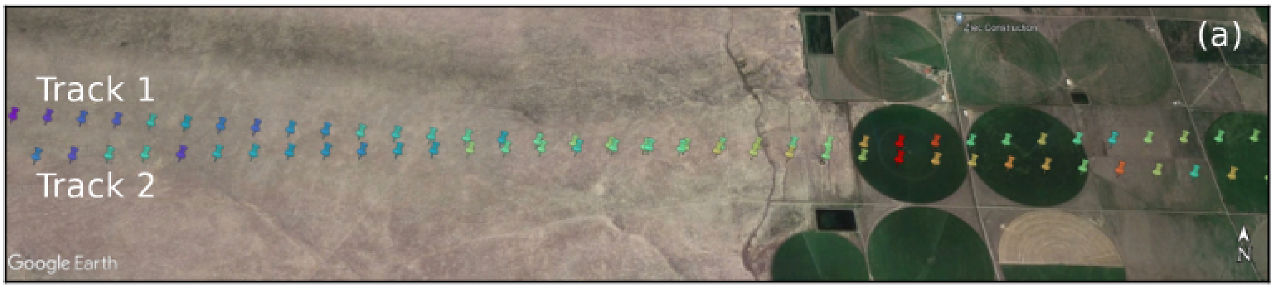


Figure 4.2.3d*: HSRR obtained in two CYGNSS tracks across San Luis Valley in nearly coincident tracks and within two weeks difference. Features at scales much finer than the low sampling rate reflectivity are consistent between the two tracks. The low sampling rate reflectivity, not shown, only provided three points along each of these segments. From [P25].*

4.2.4 Roadmap for future improvements

Using more raw IF data sets from TDS1 and CyGNSS can permit better validation and tuning the optimal integration parameters of this approach (primary coherent integration and M_c). A dedicated campaign with DoT-1 data, should it be upgraded to operate a ‘coherent channel’ could be of invaluable help.

The L1B-CC algorithm will be applied to other frequency-polarization combinations when available, and the thresholds to determine ‘coherence’ modified if needed.

5 L2OP-SI: Inundation processor

5.1 Background and scientific justification

At the kick off of this contract, there was no literature on water flagging with GNSS-R measurements, although different studies linked the detection of coherence (or high reflectivity or other indicators) to presence of water on the land surface at dedicated areas or flooding episodes. Therefore, there was not a global algorithm to discern those cases where GNSS-R presents high coherence without surface water from those that correspond to inundation, or those with low coherence over water from those over dry land.

After attempting simple algorithms for classification based on the coherence flags, we realized that the behavior in different geographic zones is important, and that many parameters might influence the final classification. At that point we investigated the possibilities of machine learning approaches and selected a Random Forest approach.

In general, the difficulty of this algorithm is the lack of actual 'ground truth' data set, as most of the surface water data bases have been generated with satellite observations that present limitations: they mostly work at electromagnetic frequencies that get obstructed by vegetation canopies and/or by clouds and precipitation, or rely on double-bounce scattering to detect water.

5.2 Algorithm description

For the incoherent channel, a Random Forest approach [L22] has been trained with three years of Power Spread Ratios computed with CyGNSS L1-LHCP data, using as reference data the monthly history of JRC's Global Surface Water database [P16]. Several combinations of input features and parameters have been attempted, with best results obtained with the following input ones: **PSR, Reflectivity, geographical coordinates of the specular point, land cover and roughness at the specular point**. The algorithm searches for the land cover of the specular point location from the ESA CCI Land Cover product [E17], currently available as L1B metadata. The roughness of the specular point location is from the roughness product on EarthEnv Topography [A18].

Similarly, for the coherent channel, a Random Forest approach [L22] was employed, using approximately 2000 raw IF data from both CyGNSS and TDS-1. The reference data for training was obtained from JRC's Global Surface Water (GSW) database [P16]. Different combinations of input fields and parameters were tested, and the optimal results were obtained using the following input variables: **Coherence coefficients, Normalized amplitude, geographical coordinates of the specular point, land cover information, and roughness at the specular point**. The algorithm also utilizes the ESA CCI Land Cover product [E17] to determine the land cover at the specular point location. The roughness data for the specular point location is derived from the roughness product on EarthEnv Topography [A18].

Therefore, both L2SI algorithms (for low rate power measurements and high rate complex signals) are essentially the same, only changing the GNSS-R observables being used (and obviously, with independent training and RF model generation). See the RF flowchart in Figure 5.2a.

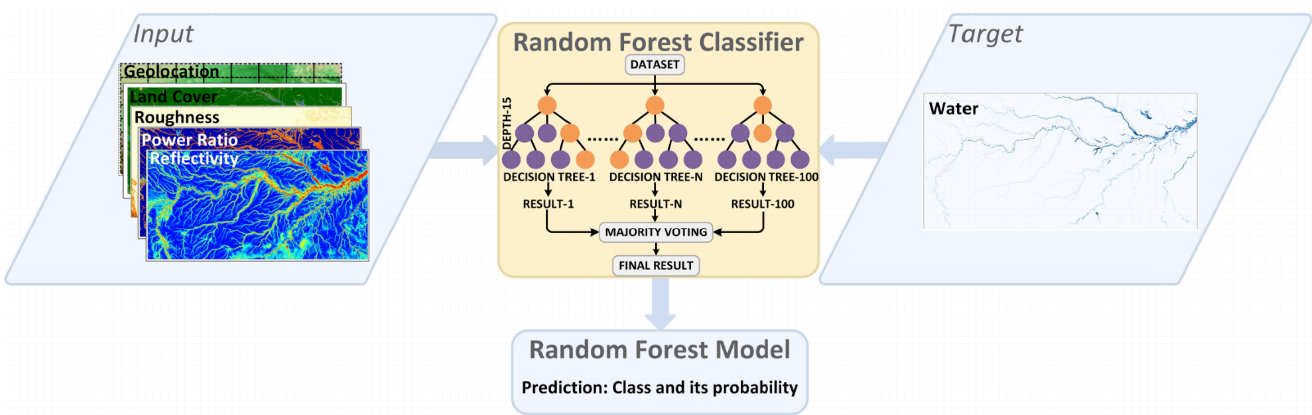


Figure 5.2a: *Flowchart for generating a data-driven water detection model using the random forest classifier. The inputs are <u>power spread ratio and reflectivity (for low-rate power measurements)</u> with geolocation, land cover and 5-km scale roughness as ancillary features, while the target is the water flag obtained from GSW [P16]. For <u>high-rate complex measurements</u>, the same approach is used, replacing the power spread ratio and the reflectivity by the <u>coherent coefficient and the normalized amplitude</u>. Figure from [P24].*

In the future, there will only be two trained RF models (and therefore, only two L2SI algorithms), as advanced in Figure 3a:

- one for low rate power measurements, which will be trained (and run) with all L1B coherent products derived from power DDMs at L1-LHCP, L1-RHCP, L5/E5-LHCP and L5/E5-RHCP reflectivities and PSRs, simultaneously; and
- another for high rate complex data, which will be trained (and run) with all L1B coherent products derived from the coherent channel at L1-LHCP, L1-RHCP, L5/E5-LHCP and L5/E5-RHCP coherence coefficients and normalized amplitudes, simultaneously.

The quality flags flow down from L1B data, while the algorithm itself provides an indicator of confidence for each of the results. See the particularities of the L2SI for low rate power measurements (L2-POW) and those of the L2SI for high rate complex data (L2-CC) in the following Sections.

It is important to highlight that the L2SI processor does not check the lower level quality flags (L1, L1B), but generates the L2SI products regardless of these flag values. It is left to the user to check the flags and decide whether to use or not the L2SI products.

5.2.1 L2-POW: Training of the Random Forest for Incoherent channel

The reference data sets for training the Random Forest algorithm is based on JRC's Global Surface Water 'Monthly History' product [P16]. This product correspond to a grid of 30 m x 30 m cells, filled with three possible values:

GSW-MH = 0 → no data corresponding to that month and cell
GSW-MH = 1 → water detected during that month and cell
GSW-MH = 2 → no water detected during that month and cell

The fine resolution of 30 m x 30 m per cell is a problem, as it is too fine compared to the HydroGNSS resolution. Given that the 'coherence' components have resolutions of up to a few km, the first step is to assign the GSW information corresponding to 3 km x 3 km area around the specular point of each CyGNSS specular point involved in the training process. This is done as described in Figure 5.2.1a, in two steps:

1. Computing the fraction of cells with water (compared to the total number of cells with GSW data) within cells of 1 km x 1 km size (see middle panel in Figure 5.2.1a).
2. Averaging the values of the 9 (1 km x 1 km) fraction-water cells around the specular point cell, accounting for the distance between the specular point location and each of the centers of the 9 cells (see right panel in Figure 5.2.1a). We call this new product 'GSfW-3km' hereafter ('f' for fraction).

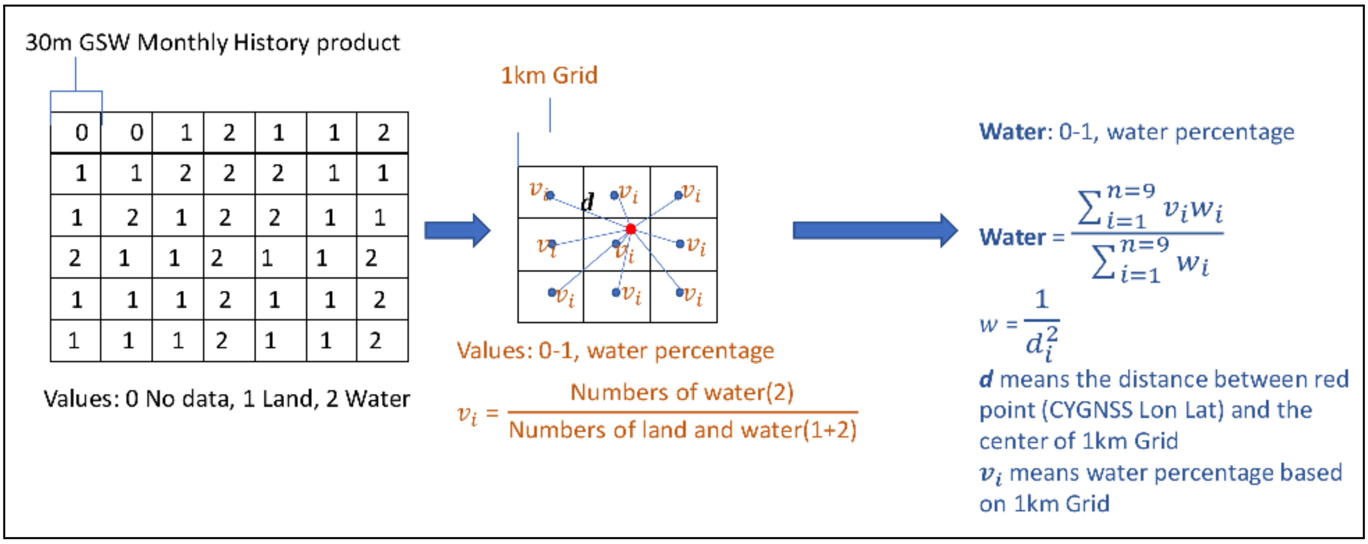


Figure 5.2.1a: *Process to up-scale the resolution of the GSW monthly history data, from 30 m x 30 m cells to 3 km x 3 km areas around each specular point involved in the training, in the form of fraction of water on the cell (GSfW-3km).*

After these steps, each specular point has a fraction of water associated with its surrounding 9 square km. Therefore, a threshold must be selected to label a 9-km² cell as 'water' or 'no water'. The threshold is set to GSfW-3km = 0.2.

Table 5.2.1a: *Threshold applied to the GSfW-3km to consider (label) the corresponding specular point over water or no water.*

Reference data over specular point as 'water'	GSfW-3km > 0.2
Reference data over specular point as 'no water'	GSfW-3km <= 0.2

Warning: the GSW products are based on images taken by Landsat 5, 7 and 8. These images are obtained in the spectral range from optical to far infrared frequencies. These bands might lose information about the surface when this is covered by vegetation. Therefore, it is expected that the reference data misses inundated areas where GNSS-R could potentially sense them (see Appendix B).

For computer memory issues, the training is split in five regions. It is important to remember that this training is limited to the zones with CyGNSS coverage, more restricted than the coverage of HydroGNSS.

The result of the training over each zone is the 'Random Forest Model', the tool used afterwards for prediction of water/no-water based on PSR and reflectivity values.

A total number of 952,027,220 CyGNSS observations were used, covering the period from January 2018 to December 2020 and excluding those specular points laying in areas with no GSW data. The training algorithm was set to work on 100 forests, 15 layers deep each. The best results were obtained when the input fields were PSR, reflectivity, land cover, roughness and GSfW-3km (in addition to the coordinates of the specular point). These input parameters are summarized in Figure 5.2.1b.

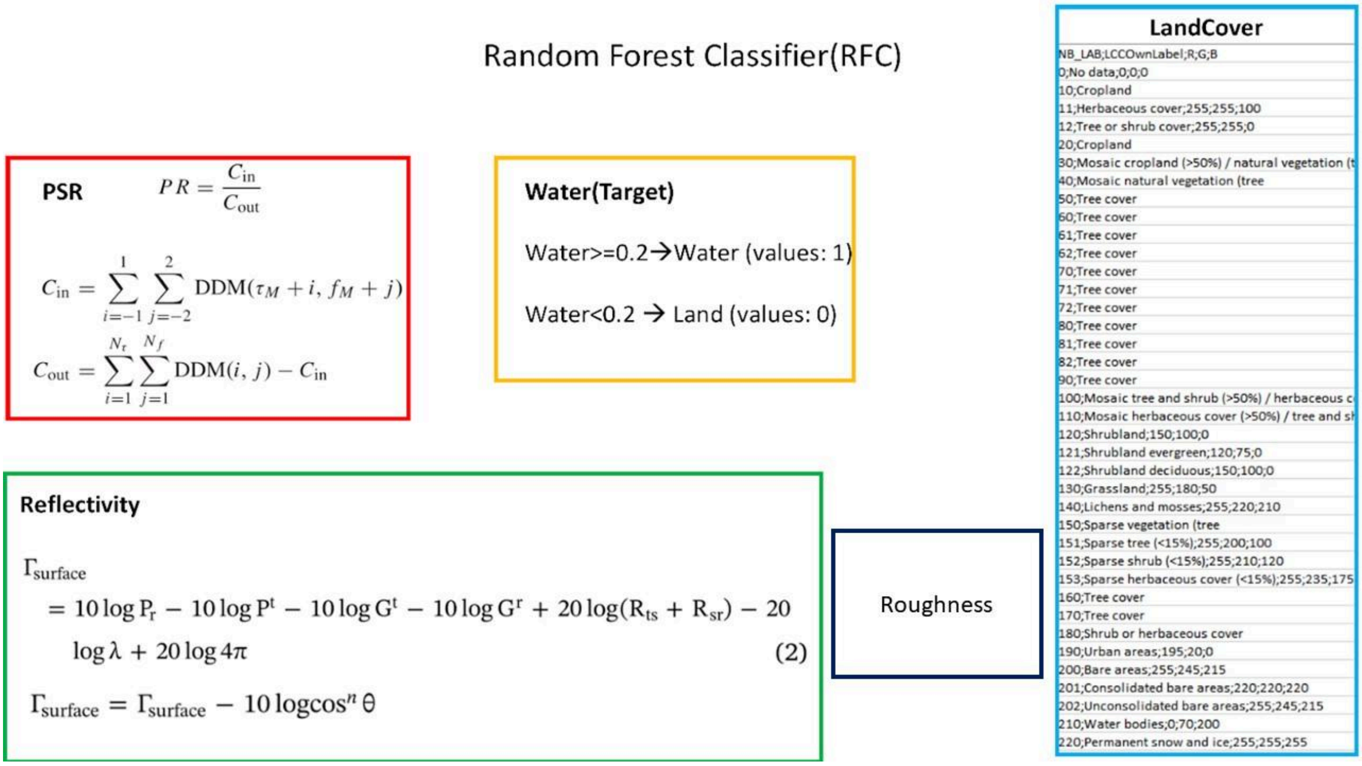


Figure 5.2.1b: *Summary of the input features and parameters used to train the random forest classifier.*

Note that the user only needs to provide the path to input L1B files, where to extract PSR, reflectivity and coordinates of the specular point. The algorithm internally searches for the land cover and roughness information of the specular point.

The random forest classifier provides values between 0 and 1 to express the confidence of the algorithm that the measurement corresponds to water. It can be understood as the probability of presence of water in each prediction. A threshold of Water Probability Incoherent >= 0.5 has been set to determine if the reflection is from a water surface, as described in Table 5.2.1b. This 'Water flag Incoherent' is also an output of the algorithm.

Table 5.2.1b: *Thresholds applied to flag a power DDM observation as 'water'.*

Coherence Flag	
No Water(Water flag Incoherent = 0)	**Water(Water flag Incoherent = 1):**
Water Probability Incoherent < 0.5	Water Probability Incoherent >= 0.5

5.2.2 L2-CC: Training of the Random Forest for Coherent Channel

The reference data sets for training the Random Forest algorithm is based on JRC's Global Surface Water 'Monthly History' product [P16]. This product correspond to a grid of 30 m x 30 m cells, filled with three possible values:

GSW-MH = 0 → no data corresponding to that month and cell (pixel without valid data)
GSW-MH = 1 → water detected during that month and cell
GSW-MH = 2 → no water detected during that month and cell

The fine resolution of 30 m x 30 m per cell is a problem, as it is too fine compared to the HydroGNSS resolution. Given that the coherence channel has resolutions of up to a few hundreds meters, the first step is to assign the GSW information corresponding to ~300 m x 300 m area around the specular point of each CyGNSS specular point involved in the training process. The GSfW-300m index is simply computed as the average of 10 GSW x 10 GSW pixels (each pixel has 30m resolution) with water data (that is, GSW-MH > 0) around the location of each specular point. If A_i^{300m} is the area of the GSfW-300m around the i-th specular point, its value is computed as the fraction between the number of 30m pixels with water detected (value 1) within A_i^{300m} divided with the total number of pixels with valid data (value > 0):

$$GSfW-300m_i = \frac{N_{(GSW-MH\ \in\ A_i^{300m})\ ==\ 1}}{N_{(GSW-MH\ \in\ A_i^{300m})\ >\ 0}}$$

A threshold must be selected to label a 300mx300m cell as 'water' or 'no water'. The threshold is set to GSfW-300m = 0.2.

Table 5.2.2a: *Threshold applied to the GSfW-300m to consider (label) the corresponding specular point over water or no water.*

Reference data over specular point as 'water'	GSfW-300m > 0.2
Reference data over specular point as 'no water'	GSfW-300m <= 0.2

Warning: the GSW products are based on images taken by Landsat 5, 7 and 8. These images are obtained in the spectral range from optical to far infrared frequencies. These bands might lose information about the surface when this is covered by vegetation. Therefore, it is expected that the reference data misses inundated areas where GNSS-R could potentially sense them (see Appendix B).

The result of the training is the 'Random Forest Model', the tool used afterwards for prediction of water/no-water based on coherence coefficients and normalized amplitude.

A total number of 2,719,720 observations from CyGNSS and TDS-1 were used, excluding those specular points laying in areas with no GSW data. The training algorithm was set to work on 100 forests, 15 layers deep each. The best results were obtained when the input fields were coherence coefficients, normalized amplitude, land cover, roughness and GSfW-300m (in addition to the coordinates of the specular point). These input parameters are summarized in Figure 5.2.2a.

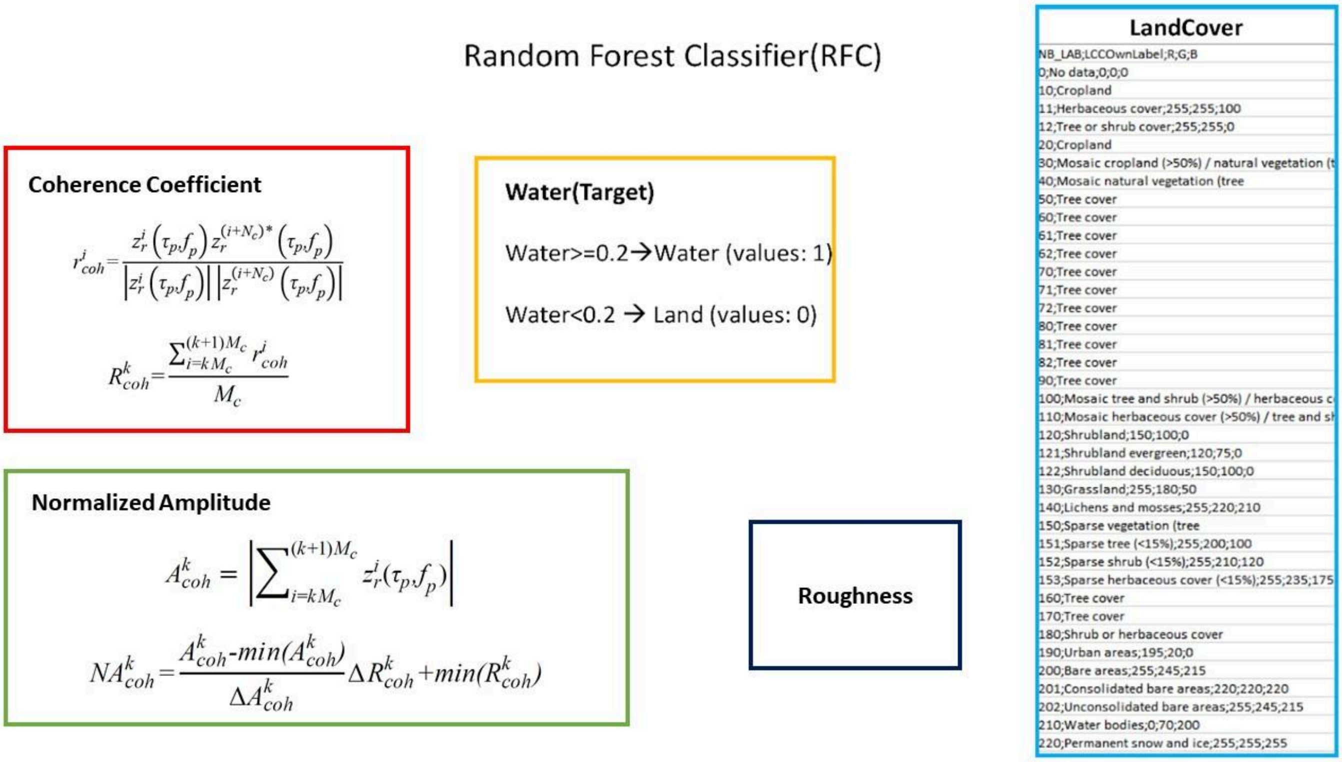


Figure 5.2.2a: *Summary of the input parameters used to train the random forest classifier.*

Note that the user only needs to provide the path to input L1B files, where to extract coherence coefficients, normalized amplitude and coordinates of the specular point. The algorithm internally searches for the land cover and roughness information of the specular point.

The random forest classifier provides values between 0 and 1 to express the confidence of the algorithm that the measurement corresponds to water. It can be understood as the probability of presence of water in each prediction. A threshold of Water Probability Coherent >= 0.5 has been set to determine if the reflection is from a water surface, as described in Table 5.2.2b. This 'Water flag Coherent' is also an output of the algorithm.

Table 5.2.2b: *Thresholds applied to flag a power DDM observation as 'water'.*

Coherence Flag	
No Water(Water flag Coherent = 0)	**Water(Water flag Coherent = 1):**
Water Probability Coherent < 0.5	Water Probability Coherent >= 0.5

5.2.3 Interface of L2OP-SI (Beta version)

***Table 5.2.3**: Interface of the L2OP-SI processor algorithm. The user only needs to call it after changing the input path in the configuration file, but the inner algorithm is called with particular inputs and outputs as described below.*

<table>
<tr><th colspan="2">Interface of L2OP-SI beta version</th></tr>
<tr><td>SYNTAX (full directory):</td><td>L2OP_SI_DR.exe</td></tr>
<tr><td>INPUT</td><td>OUTPUT</td></tr>
<tr><td>- Power Spread Ratio (PSR) @ T_{incoh} rate
- Reflectivity @ T_{incoh} rate
- L1B Quality Flags
- Specular Point geographic coordinates
- LandType L1B metadata (from ESA CCI)

- Coherence Coefficients (CC) @ T_{coh} rate
- Normalized Amplitude (NA) @ T_{coh} rate

- L1B Quality Flags

Databases:
- Random Forest Trained Model
- LandType from ESA CCI
- Roughness from EarthENV Topography</td><td><u>group: Low Resolution → Low sampling rate (@ T_{incoh}):</u>

-‘WaterFlagIncoherent’: Water Flag at @ T_{incoh} rate based on Water Probability. 1=water, 0=no water. [16-bit unsigned integer]

-‘WaterProbabilityIncoherent’: Confidence of the Random Forest algorithm that the result for that observation corresponds to water. It can be understood as probability predicted by the algorithm to be water. This is provided at @ T_{incoh} rate based on PSR and reflectivity [32-bit floating point]

- ’DDMPeakOffsetDistance’: as defined in L1B (passed from L1B), is considered poor quality when it exceeds 7.

- ’DDMPeakQualityFlags’: as defined in L1B (passed from L1B)

- ‘SIOverallQualityFlags’: Low quality summary flag composed of a bitwise or of 'LowOverallQuality' (flag from others L1 algorithms), 'DDMPeakQualityFlags', 'DDMPeakOffsetDistance'. Flag that takes values 1 (Bad) and 0 (Good). This is provided @ T_{incoh} rate. [16-bit unsigned integer]

- ‘IntegrationMidPointTime’: time at the half-window of the observation.
- ‘SpecularPointLon’: longitude of the coordinates of the specular point at IntegrationMidPointTime’.

- ‘SpecularPointLat’: latitude of the coordinates of the specular point at IntegrationMidPointTime’.

<u>group: High Resolution → High sampling rate (@ T_{coh}):</u></td></tr>
</table>

	- ‘**Time_CoherentInt**’: Time when the coherent measurement took place. Provided @ T_{coh} rate. - ‘**Longitudes_CoherentInt**’: Longitude of the coordinates of the specular point of the coherent result (‘Time_CoherentInt’). Provided @ T_{coh} rate. - **‘Latitudes_CoherentInt’**: Latitude of the coordinates of the specular point of the coherent result (‘Time_CoherentInt’). Provided @ T_{coh} rate. -‘**WaterFlagCoherent**’: Water Flag @ T_{coh} rate based on Water Probability. 1=water, 0=no-water. [16-bit unsigned integer] -‘**WaterProbabilityCoherent**’: Confidence of the Random Forest algorithm that the result for that observation corresponds to water. It can be understood as probability predicted by the algorithm to be water. This is provided at @ T_{coh} rate based on coherence coefficients and normalized amplitude [32-bit floating point]
SAMPLING RATE:	Low sampling rate (@ T_{incoh} rate, baseline 1 or 2 Hz) and high sampling rate (@ T_{coh} rate, baseline 25 Hz).
TYPE:	Along track product, no gridding, no interpolation.

5.3 Validation

Appendix B shows validation results obtained with the L2OP-SI applied to CyGNSS data.

5.4 Roadmap for future improvements

The training of the L2-CC with a much larger set of data could improve the algorithm.

Finding proper ‘surface water’ reference datasets could help improve the algorithm. A better reference would also be good for proper validation of the algorithm.

Once data at dual-polarization and dual-frequency is obtained, both low-rate and high-rate algorithms should be re-trained fed by all frequency-polarization combinations simultaneously. Initial attempts could be done with airborne data, if available.

6 Spatial Resolutions

Table 6 summarizes the different spatial resolutions together with the rationale used to estimate them.

Table 6: *Spatial resolutions corresponding to different coherence and water flagging products.*

Spatial Resolutions		
Product:	**Spatial Resolution:**	**Rationale:**
PSR	~20 km	The PSR algorithm uses an extended area within the DDM, therefore, the displacement of the specular point during T_{incoh} is exceeded by the area from which the box of delay-Doppler cells is representative. This is estimated at ~20 km when using a few cells box. However, when the coherent component of the scattering is dominant, the reflected signals might come from a much reduced area, down to e.g., ~500 m (approximate size of the First Fresnel Zone in near-nadir geometries).
Pcc	~7 km	The Pcc algorithm is based on the coherent coefficients, which contain information about a limited area (a few hundreds of meters). As the information is accumulated by T_{incoh}, the spatial resolution is mostly limited by the spacing of T_{incoh} (~7 km if 1 second incoherent sampling is used).
R_{coh}	~1 km	The algorithm R_{coh} is based on the coherent coefficients, which contain information about a limited area (a few hundreds of meters). The coherence coefficients will be further coherently integrated (factor M_c), degrading the spatial resolution. If the nominal T_{coh} is 4 ms, the displacement between samples is ~30 meter, so M_c=10 samples integration would correspond to a displacement of ~300 meter, meaning a slight degradation of the resolution. A conservative ~1 km is suggested here.
Water flag Incoherent	~20 km	The water flag incoherent obtained from the PSR and reflectivity products is representative of the same area as the PSR. Nevertheless, the training is done with GSW information over 9 square km, so the training process might have refined the spatial resolution.
Water flag Coherent	~1 km	The water flag coherent obtained from the coherence coefficients and normalized amplitude products is representative of the same area as the R_{coh}.

7 Mapping algorithms with Delivered versions and Reference data

Table 7a: *Algorithms delivered or expected to be delivered in each of the versions. For those delivered, the version ID is provided*

	L1B-CX	L1B-CC	L2OP-SI	
			L2-SI PSR	L2-SI CC
Placeholder	L2PP-PSR v1		(dummy, in L2PPSI-PSR v1)	
Alpha	L1B-CX v1	L1B-CC v1	L2OPSI-PSR v2	
Beta	L1B-CX v2	L1B-CC v2	L2OP-SI (including sub-processors L2OPSI-POW v3 and L2OPSI-CC v1)	

Table 7a above summarizes the algorithms delivered or to be delivered in each version of the PDGS and E2ES.

The placeholder version of the delivered code was called L2PPSI and included the L1B-CX. However, the water flag was simply a copy of the coherence flag.

At the alpha version, the L1B-CX algorithm has been delivered separately, so it can also run to extend L1B products, independently of the L2 SI. The L1B-CC algorithm was also implemented and delivered independently. Finally, the alpha L2OP-SI algorithm included a proper L2 SI random forest algorithm for classification of presence of water, based on power measurements (output of L1B-CX).

For the Beta version, we expand the L2-SI algorithm to include the coherence channel L1B-CC product for a higher resolution water flag. The time frame does not seem sufficient to implement a L2 algorithm that uses both power and complex indicators, so we envisage two different products under high and low resolution groups in the same output file.

The L2OP-SI algorithms require reference data for training purposes. Table 7b links each processor version with the version of the Random Forest Model generated (and used by the processor) together with the reference data of the training set.

Table 7b: *Link between the L2OP-SI algorithm version ID, the Random Forest Model version ID used by the processor, and the reference data used to train it.*

L2OP-SI ID/version:	Random Forest Model ID/version:	Reference data for training:

L2OPSI-POW v3 *(beta)*	RFM v3	CyGNSS L1 Science Record v2.1 (2018) CyGNSS L1 Science Record v3.0 (2019-2020) JRC's Global Surface Water 'Monthly Water History' product [P16], VER4-0 (2018-2020)
L2OPSI-CC v1 (beta)	RFM v1	All available CyGNSS and TDS-1 Raw_IF datasets until Jan 2022 and corresponding Global Surface Water 'Monthly Water History' product

Appendix A: INTERNAL TESTS FOR COHERENCE L1B PROCESSOR

Synthetic data has been generated in three different regimes:

- Purely coherent tones with some residual Doppler effect. They all have amplitude 1 and a phase rotating at the residual frequency.
- Purely random phasors. In this case, both the amplitude and the phase take random values.
- Mixed time series. To generate these series, a pure tone of amplitude 1 is added to a series of random phasor. The random series can have different mean amplitude, changing the overall signal to noise ratio (SNR). The SNR is then computed as the ratio of the tone's amplitude (1) to the mean of the random phasors' amplitudes, expressed in dB.

The following pages contain sets of plots, grouped by the SNR. Each row of plots shows:

- Left: 1 second series of the complex phasor at 4 msec sampling, in the complex space.
- 2nd-Left: The time series of the phases of these phasors.
- 3rd-Left: The series of complex r_coh in the complex space.
- Right: The time series of the resulting coherence coefficient R_coh

The colors of the plots are the same as in Figure 4.2.2b.

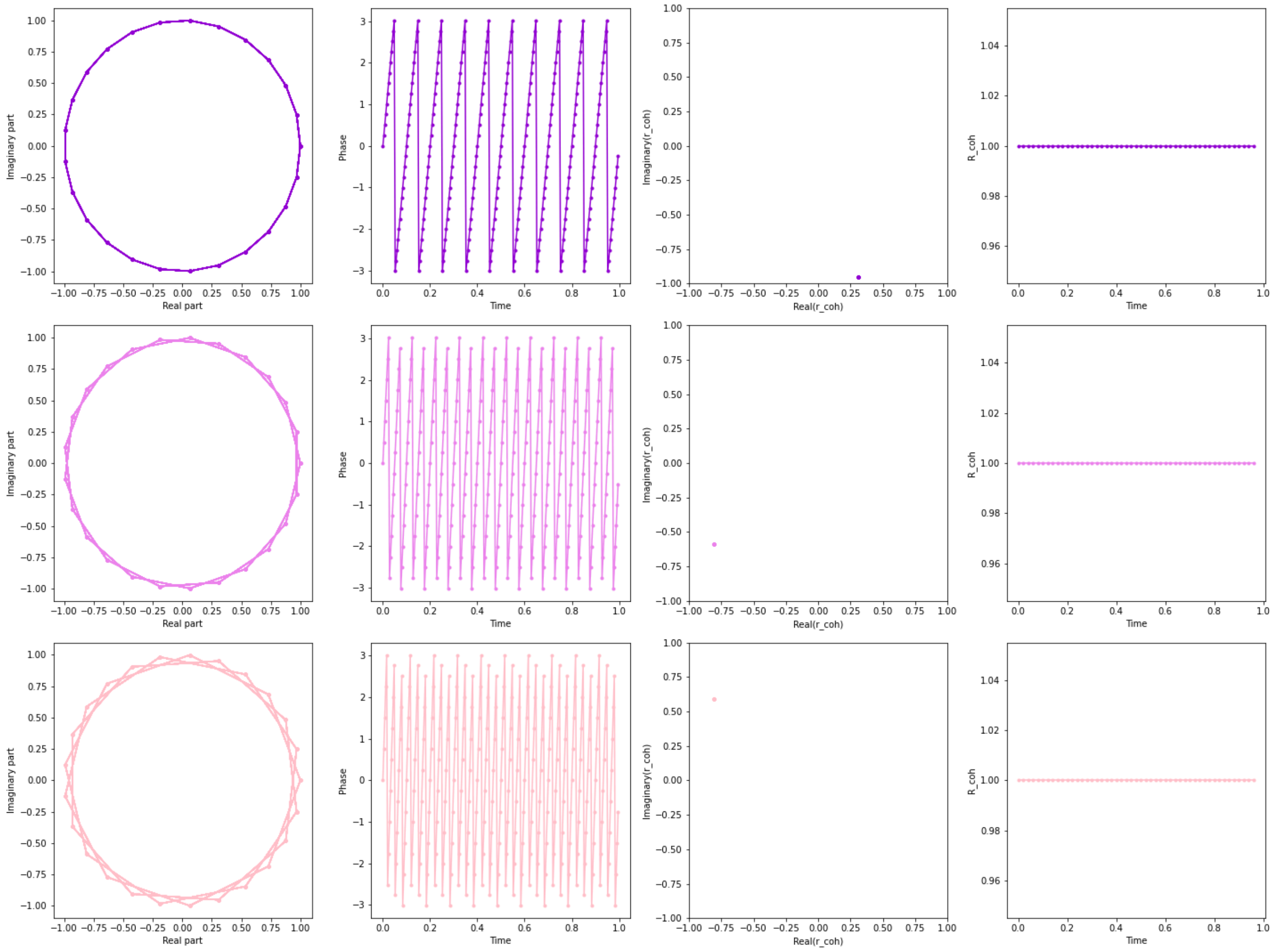


***Figure A1**: Plots for 1 second of pure tones, each with a different Doppler residual. From top to bottom: 10Hz residual Doppler, 20Hz and 30Hz. Left: 1-second series of the complex phasor at 4 msec sampling, in the complex space. 2nd-Left: The time series of the phases of these phasors. 3rd-Left: The series of complex r_{coh} in the complex space. Right: The time series of the resulting coherence coefficient R_{coh}.*

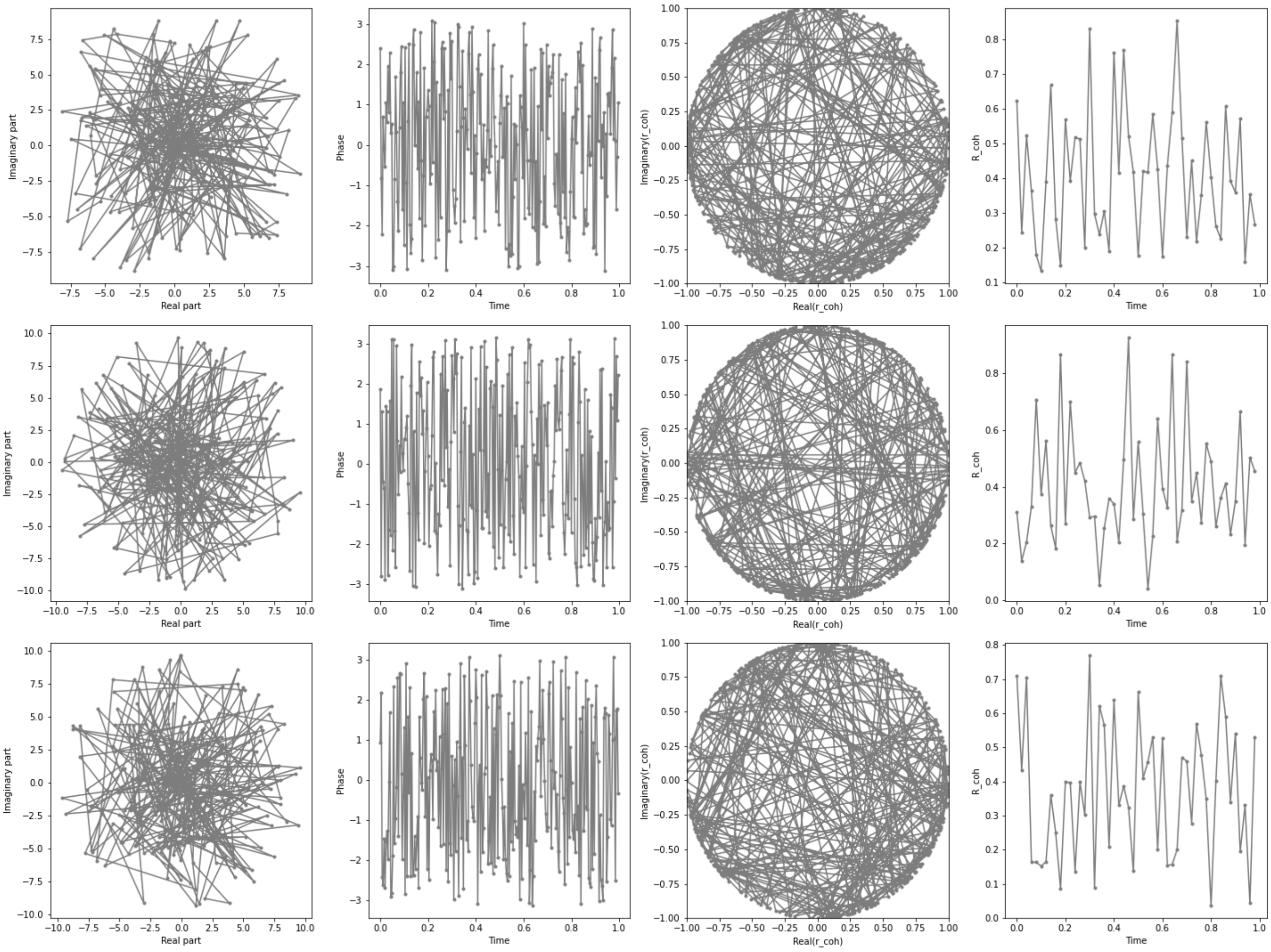


Figure A2: *Same as A1 for 1 second of pure random phasors. Three different realizations.*

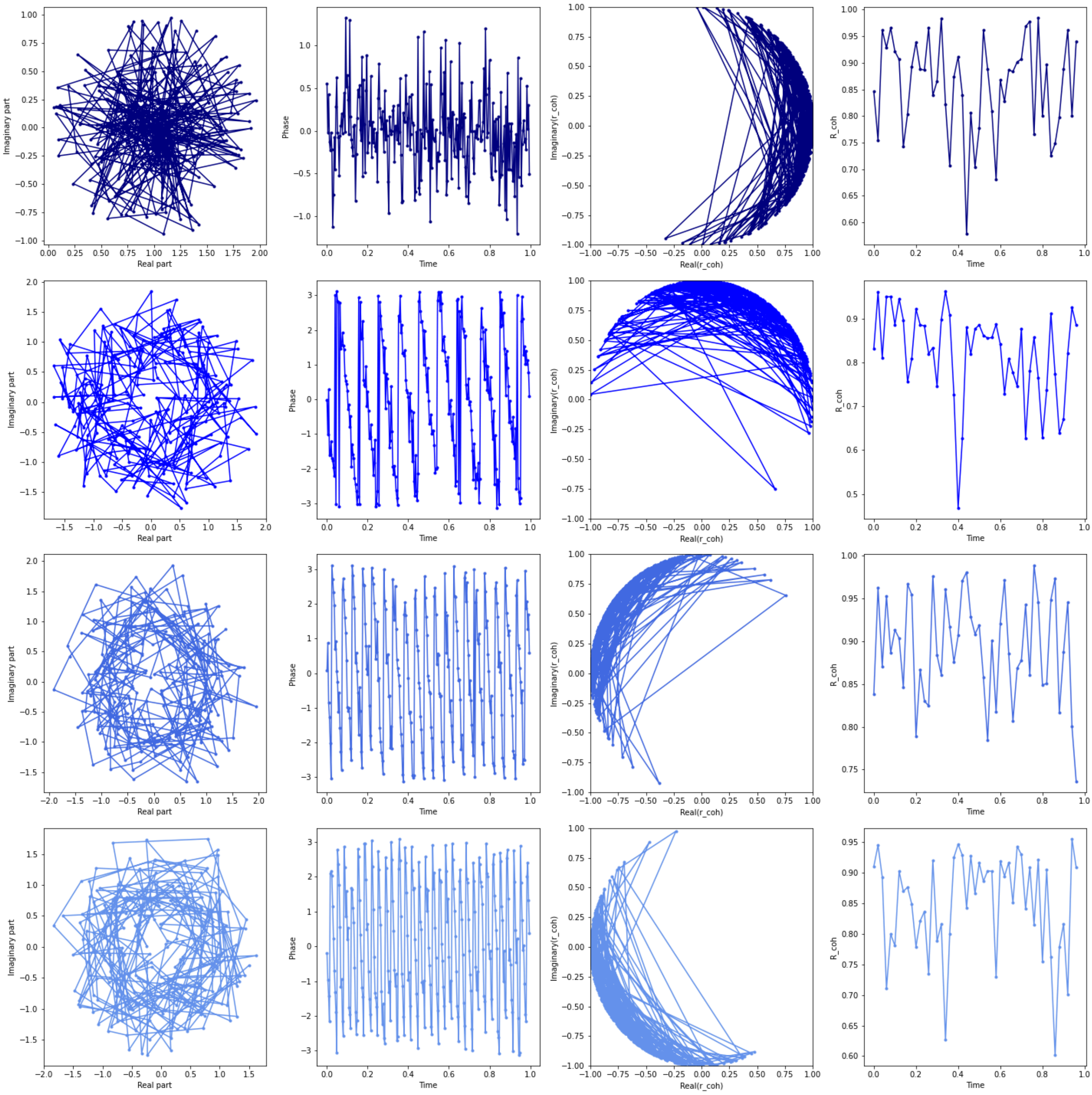


***Figure A3**: Same as A1 for 1 second of data computed with mixed (tone+random) phasors. From top to bottom: Doppler=0Hz, SNR=6.6dB; Doppler=10Hz, SNR=6.3dB; Doppler=20Hz, SNR=5.6dB; and Doppler=30Hz, SNR=6dB. Left: 1-second series of the complex phasor at 4 msec sampling, in the complex space. 2nd-Left: The time series of the phases of these phasors. 3rd-Left: The series of complex r_{coh} in the complex space. Right: The time series of the resulting coherence coefficient R_{coh}.*
The visual inspection suggests that these are all coherent, the phases could be tracked at least for fractions of the second.

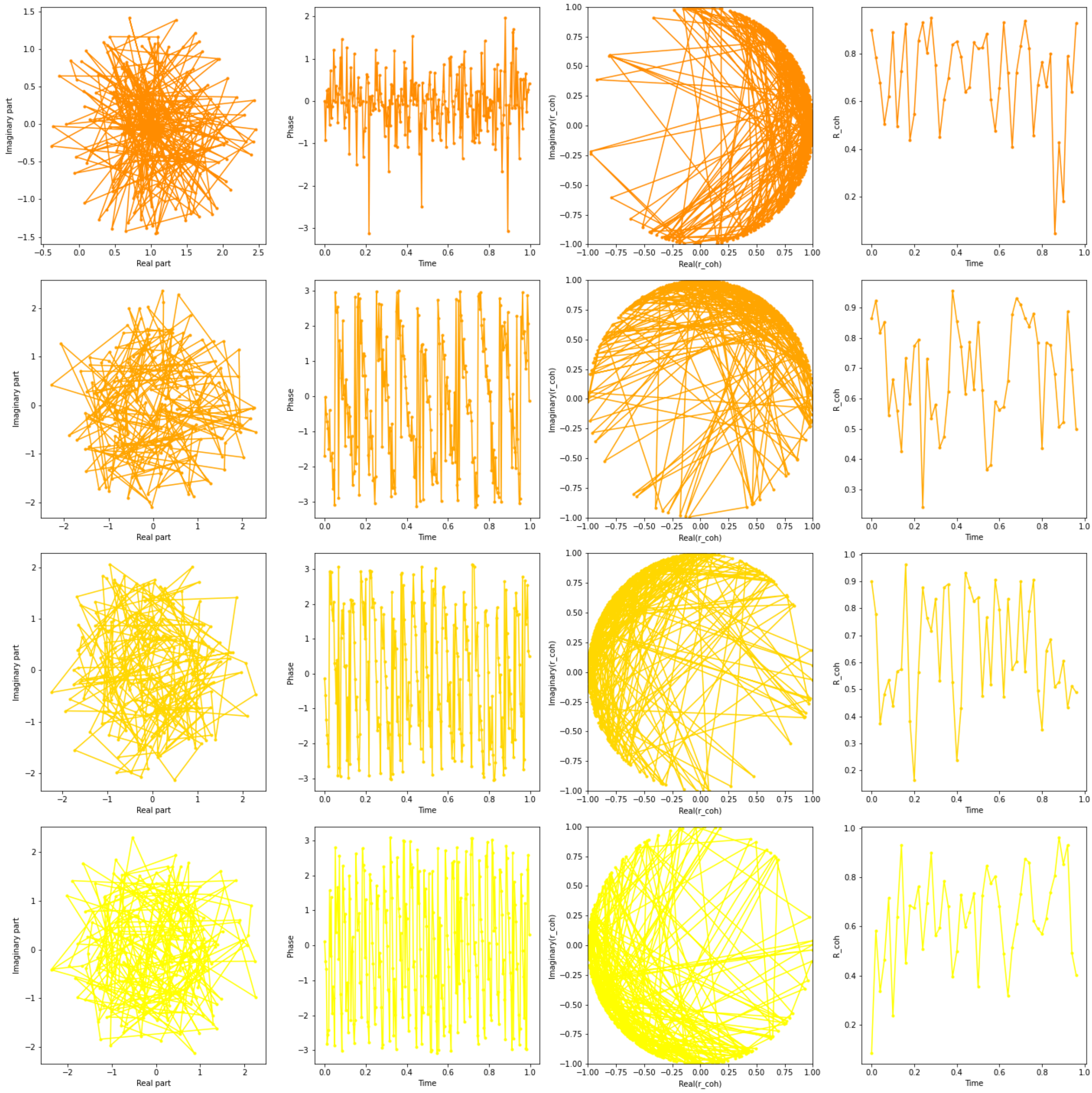


Figure A4: *Same as Figure A3 for (top to bottom): Doppler=0Hz, SNR=2.1dB; Doppler=10Hz, SNR=2.4dB; Doppler=20Hz, SNR=2.5dB; Doppler=30Hz, SNR=2.5dB. Left: 1-second series of the complex phasor at 4 msec sampling, in the complex space. 2nd-Left: The time series of the phases of these phasors. 3rd-Left: The series of complex r_{coh} in the complex space. Right: The time series of the resulting coherence coefficient R_{coh}.*

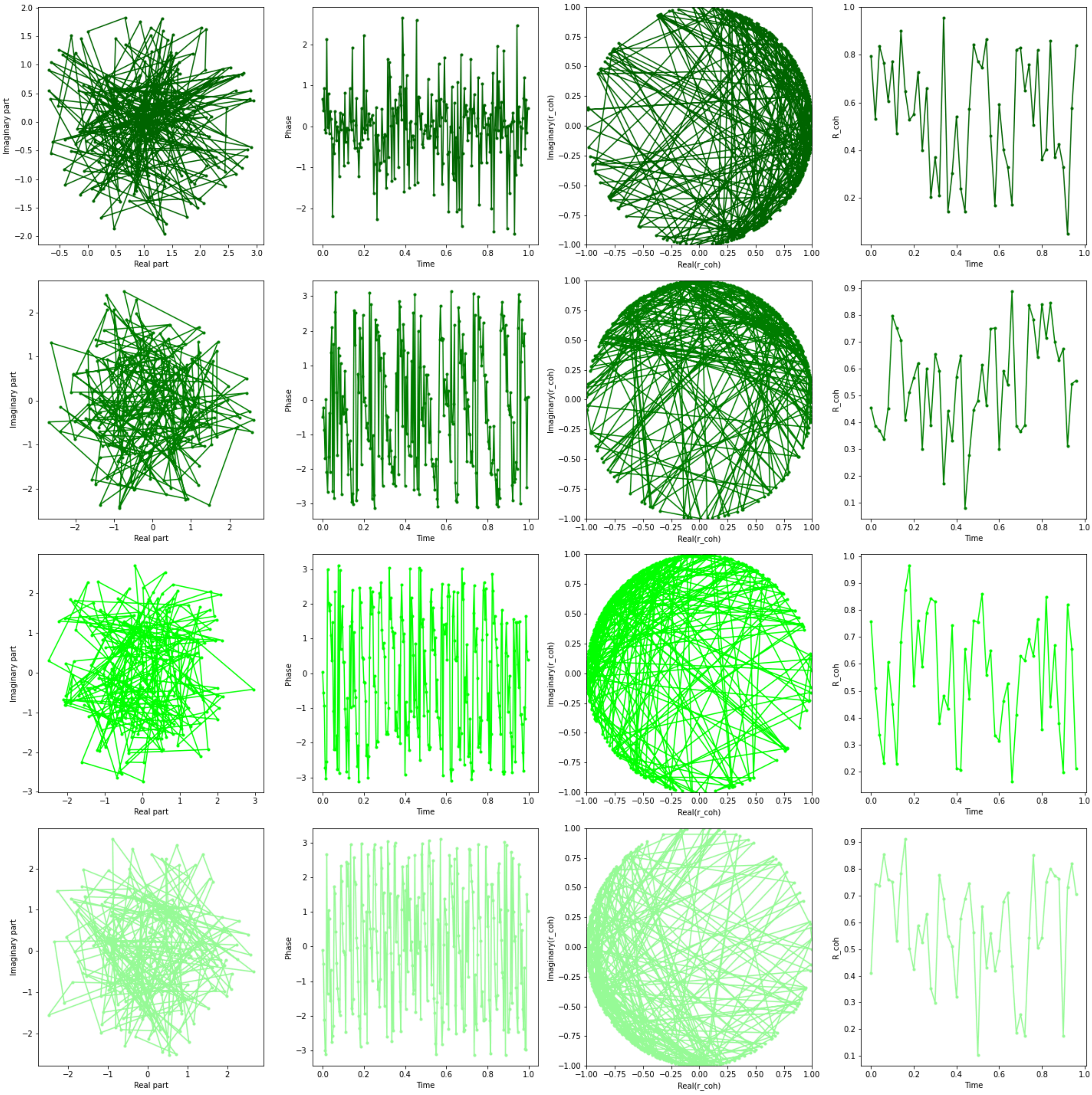


Figure A5: *Same as Figures A3 and A4, for the cases (top to bottom): Doppler=0Hz, SNR=0.3dB; Doppler=10Hz, SNR=-0.6dB; Doppler=20Hz, SNR=-0.1dB; Doppler=30Hz, SNR=0.4dB. Left: 1-second series of the complex phasor at 4 msec sampling, in the complex space. 2nd-Left: The time series of the phases of these phasors. 3rd-Left: The series of complex* r_{coh} *in the complex space. Right: The time series of the resulting coherence coefficient* R_{coh}*.*

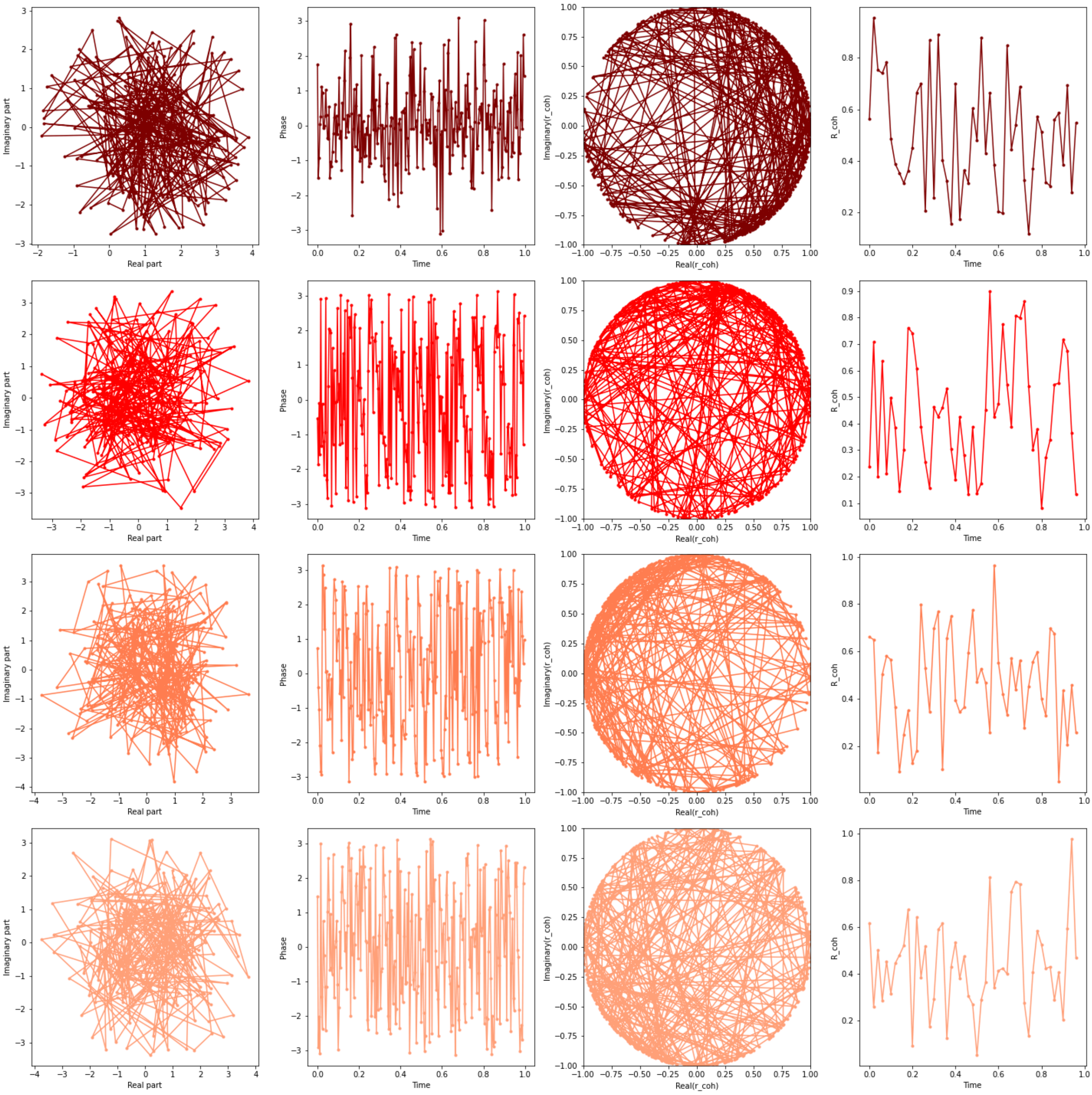


Figure A6*: Same as Figures A3 to A5, for cases (top to bottom): Doppler=0Hz, SNR=-3.7dB; Doppler=10Hz, SNR=-3.4dB; Doppler=20Hz, SNR=-3.9dB; and Doppler=30Hz, SNR=-4.1dB. Left: 1-second series of the complex phasor at 4 msec sampling, in the complex space. 2nd-Left: The time series of the phases of these phasors. 3rd-Left: The series of complex r_{coh} in the complex space. Right: The time series of the resulting coherence coefficient R_{coh}.*

Appendix B: VALIDATION OF L2OP-SI

B.1 Validation of the selected approach

The validation of the L2OP-SI (beta) algorithm from the incoherent channel was conducted for both low sampling rate power measurements and high sampling rate complex values. In all cases, only LHCP polarized signals have been used. The validation study is detailed in [P24]. Below we summarize the main aspects.

B1.1 SI from low sampling rate power observables

In this section we summarize the L2OP-SI algorithm based on slow rate power DDM information (based on observables from L1B-CX algorithm). The training uses CyGNSS data for a period of 3 years (2018 to 2020), and the testing set is from January 2021 to June 2021.

The preliminary overall global accuracies in this initial test data set is as presented in Table B1:

Table B1: *Preliminary accuracies of the predicted water flag as obtained with the L2OP-SI (beta version) with 6 months of data (Jan-Jun 2021).*

Predicted Water / True Water = 0.95	**Predicted Water / True Land = 0.05**
Predicted Land / True Water = 0.05	**Predicted Land / True Land = 0.95**

The fractions computed over Predicted values are strongly affected by the unbalanced sets. When the statistics are computed using random subsets of balanced True samples (same number of True Land and True Water samples), the results converge with those in Table B. Figure B1 shows the distribution of F1-scores when computed across 49 subsets of the data, each of the subsets selected to have a balanced data (balanced here meaning that the amount of CyGNSS measurements labeled as land by GSW has the same size as those marked as water).

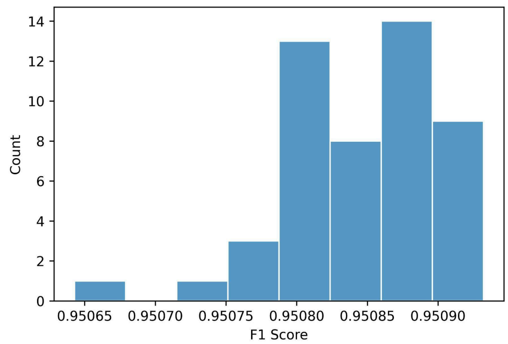


Figure B1: *F1-scores from 49 subsets of data randomly selected but forcing water/land balance.*

Initially, the ESA CCI land cover information was not used in the random forest algorithm, but this resulted in lower Correct negative scores and increased false positives. Different scenarios induce different scattering, and the land cover of the surface is an important factor affecting the response. Land cover has been shown to affect the peak reflected power, so it will be helpful to distinguish different scenarios.

While roughness affects reflectivity, flat areas typically exhibit higher reflectivity, even in an unflooded state. Therefore, it is crucial to incorporate roughness into the model to accurately identify areas with high reflectivity.

The geographical location of the specular point was neither used as input to the algorithm in its initial attempts, but the performance improves significantly when the geo-location is considered. An explanation to this might be the complexity of other effects that alter the coherence and reflectivity and are particular to a small area, such as, topography and surface roughness, vegetation, etc. To illustrate this, Figure B2 shows the distribution of PSR and reflectivity grouped by the GSW classification in three different regions. The differences in distribution shape, ranges and class separation is clearly shown and justifies feeding the random forest algorithm with the geo-location of the specular point.

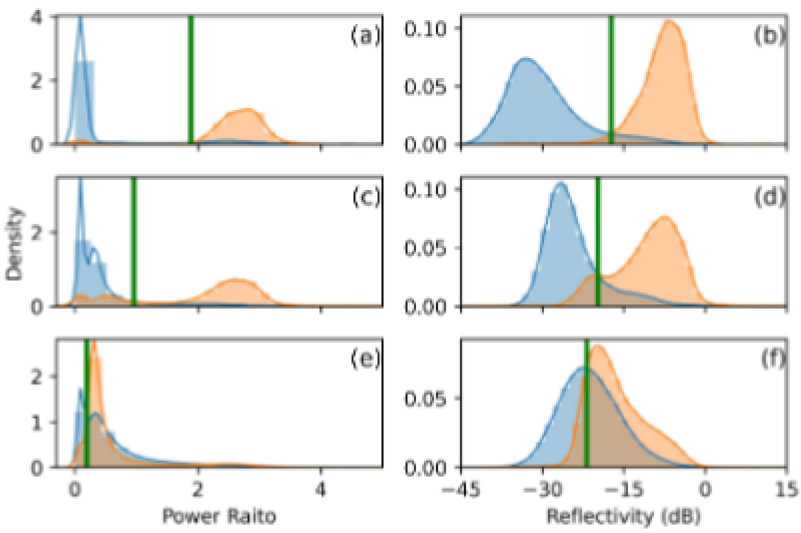


Figure B2: *Different probability density functions (PDFs) of power ratio (left) and reflectivity (right) based on three experimental areas from January 2018 to June 2021. Blue cases: the PDFs of points that are labeled as land by GSW; orange cases: the PDF of points that are labeled as water by GSW; green line: the intersection of the two PDFs. Top(a, b), middle (c,d), bottom(e,f) are regions within Congo, South America and the North shore of Lake Victoria zones, respectively. The Congo zone covers latitudes 0◦N to 1◦30'N and longitudes 23◦E to 26◦E; the zone in South America is at latitudes 20◦30'S–23◦S and longitudes 50◦W–54◦W (across Sao Paulo and Paraná states in Brazil); the Northern shore of Victoria Lake area expands into Northern Uganda, between latitudes 0◦S to 3◦S and longitudes 31◦E–35◦E.*

Figure B3 shows the locations of CyGNSS flagged results for 6 months in 2021 (Jan-Jun 2021). It generally presents the features of rivers and other well known water bodies.

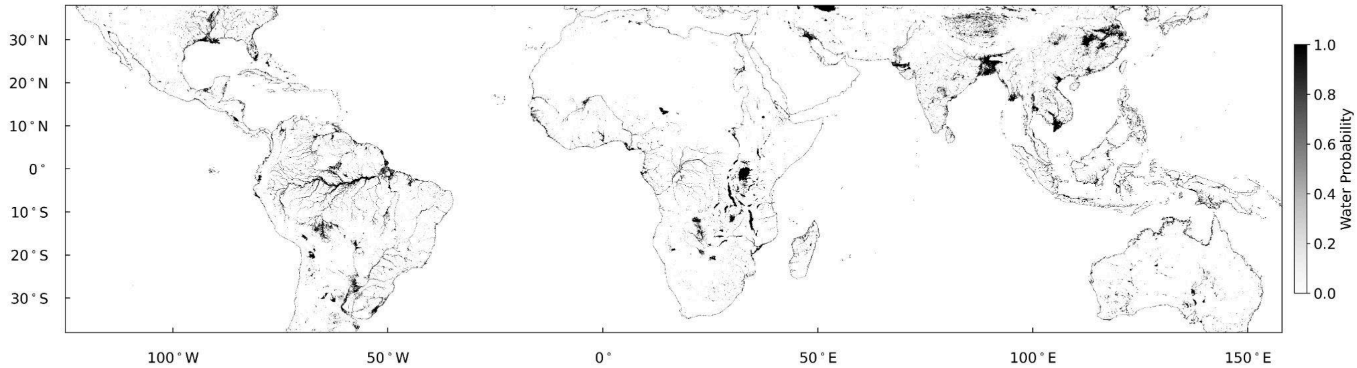


Figure B3: *The results of the L2OP-SI beta water flag for January-June 2021 CyGNSS data. The water percentage indicates the percentage of points predicted to be water in each 3.125km grid.*

Figure B4 shows the false positives for the same period. Since the number of false positives is much greater than that of false negatives, we only check the spatial distribution of false positives. It can be seen that many of these false positives seem to correspond to areas with high chances of water. Therefore, there is the possibility that may correspond to true positives, but that the reference data, GSW obtained with Landsat 5, 7 and 8, poorly sensitive below the vegetation, did not capture the water (then, false negatives in the reference data).

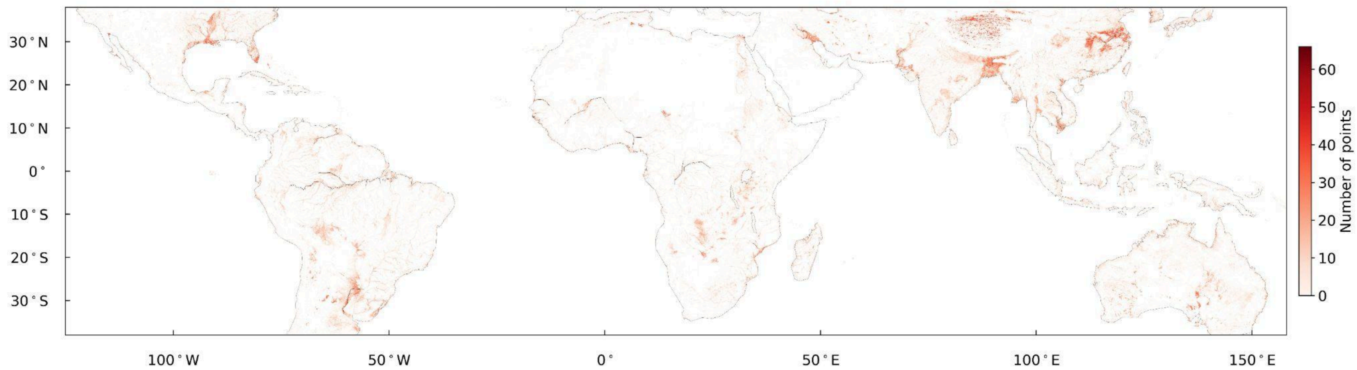


Figure B4: *L2OP-SI beta 'false positives' (according to the reference data) for the same period as in Figure B3. The number of points means how many points are false positive in each 3.125km grid.*

Figure B5 shows the probabilities for the same period. It can be seen that the probability in some regions next to rivers or wetlands and Qinghai-Tibet Plateau is lower than the water regions. It is as expected that those regions have the variability as the different degrees of surface inundation or the ambiguity of coherence indicators. And it is a good way to represent the confidence of each prediction.

To check the spatial resolution of this approach, figures B6 and B7 show a zoom into a smaller region. B6 is similar to B3, for January 2021 and over a region of 10º longitude x 10º latitude, and it is compared to the Global Surface Water for the same month. The features shown in B6 are smaller than 25km, the required resolution. Nevertheless, given that B6 has been gridded to 3.125 km cells, B7 shows the values without any further gridding on an even smaller sub-area.

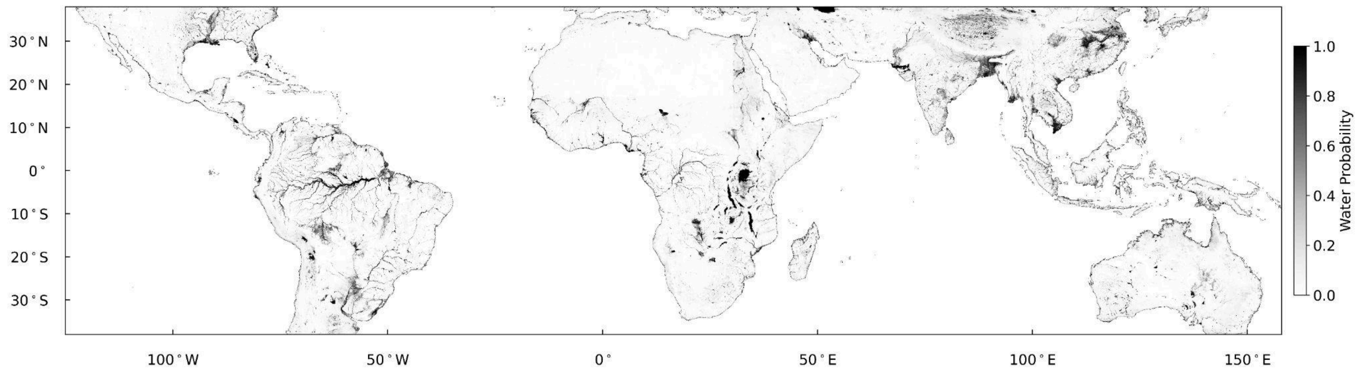


Figure B5: *The probability of the L2OP-SI beta water flag for January-June 2021 CyGNSS data. The probability represents the average of the probabilities of prediction as water in each 3.125 km grid.*

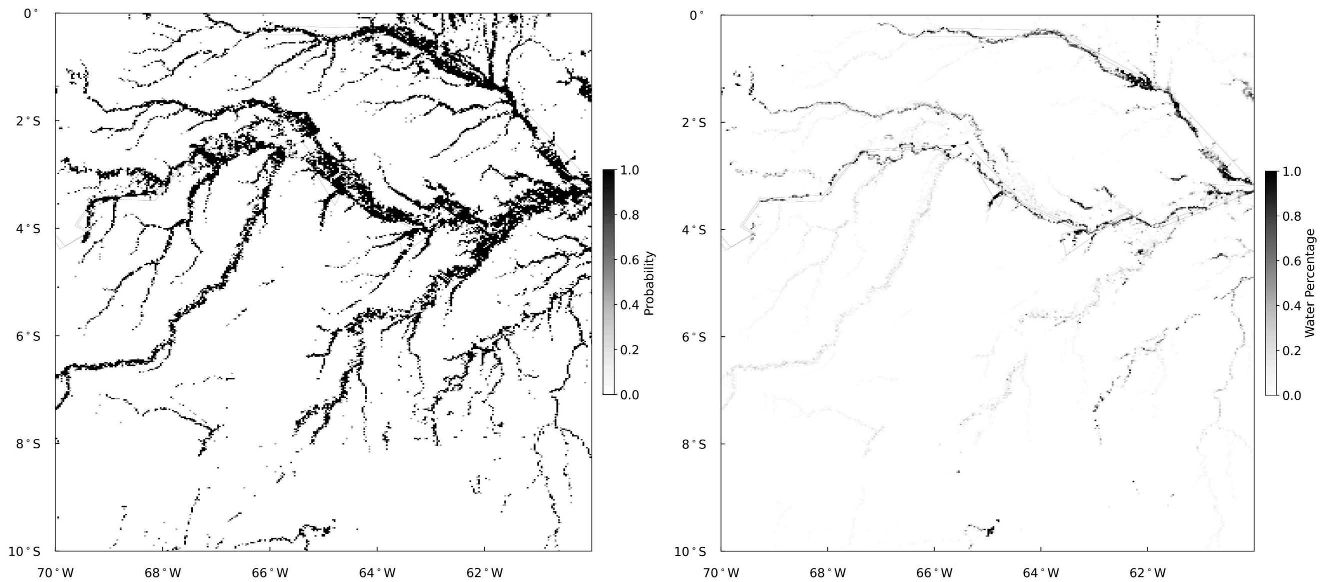


Figure B6: *(Left) CyGNSS water percentage as computed in Figure B3, here for January 2021 solely. (Right) The Global Surface Water values for the same month (Jan'21).*

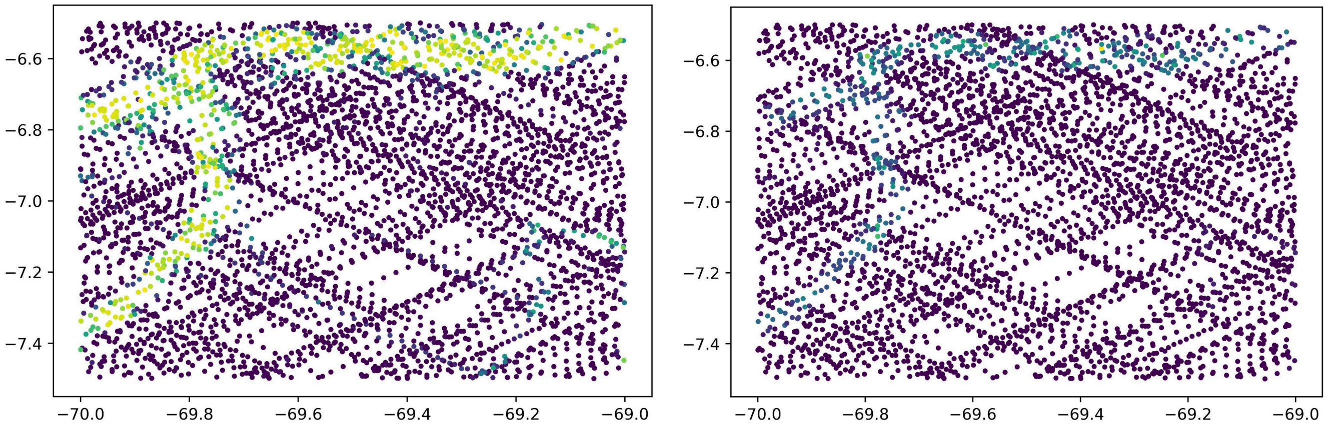


Figure B7: *A smaller sub-region of Figure B6 (1ºx1º), but now without any 3.125km gridding, that is, showing the values of 'WaterProbability' given by the L2OP-SI PSR processor as applied to CyGNSS data (left) and the co-located value as given by the Global Surface Water (right). Both correspond to January 2021. Note that the GSW has difficulties detecting water under the thick vegetation of the area (lighter blue).*

The validation of the L2OP-SI (beta) algorithm from the coherent channel is being conducted, which only includes high rate complex signals (from L1B-CC algorithm).

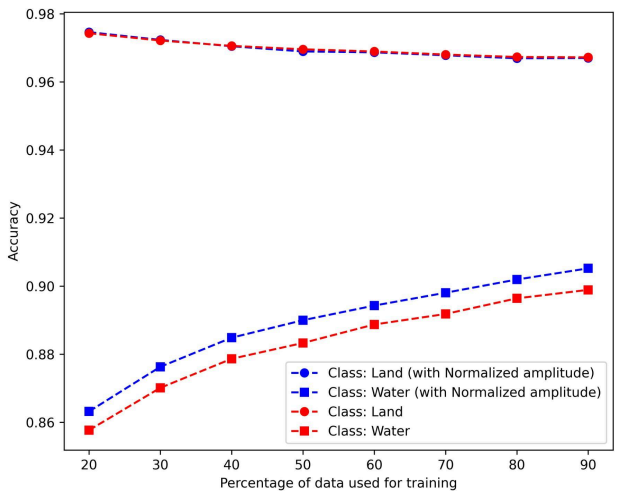


Figure B8: *The performance of different percentages of data used as the training dataset*

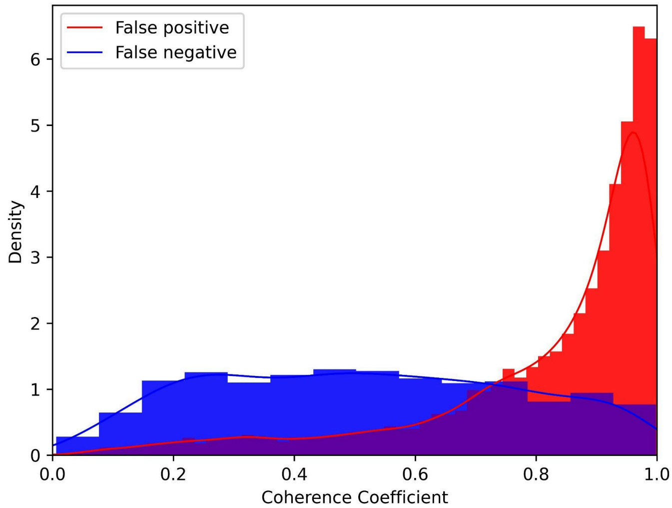


Figure B9: *The distribution of false positive and negative points for the group with 80% of the data used for training*

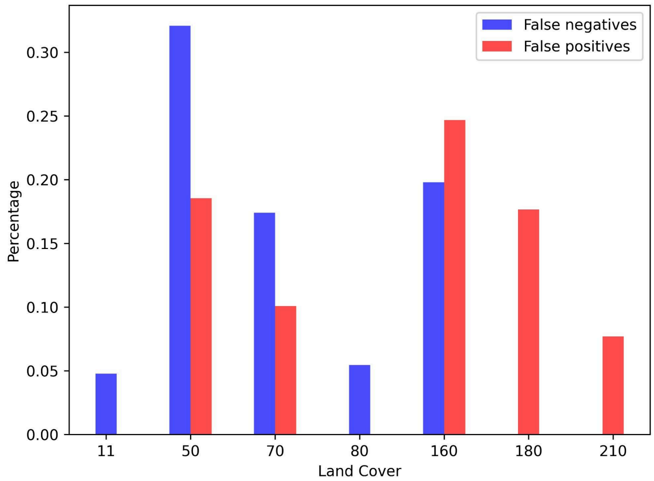


Figure B10: *The percentage of main land types on false positive and negative points for the group with 80% of the data used for training. Herbaceous cover: 11, 180; Tree cover: 50, 70, 80, 160; Water bodies: 210*

B1.2 SI from high sampling rate complex observables

As detailed in [P24, Sec.V-D], this study has been done with ~108000 seconds of raw IF data collected with both CYGNSS and TDS1 missions, as processed by in-house software receivers at ICE-CSIC/IEEC, further coherently integrated to 4 milliseconds and the single pixel complex signal of the specular point stored to mimic HydroGNSS coherent channel. The complex coherence indicators are then generated with the L1-CX algorithm. The data set was separated for training and validation as shown in Table B1.2a.

Table B1.2a: *Complex signals used for the study, given in 40 ms samples (output of L1-CX algorithm).*

Data type	Training (before Dec'21)	Validation (after Dec'21)
High rate complex field	2,175,776 samples	543,944 samples

The first concern was the consistency between CYGNSS and TDS-1 results, as they came from missions and receivers with different specifications, trained blindly by the algorithm, irrespective of these differences. Both datasets present very similar distributions of CCs and NAs, an indication that the different origin of the signals is not an issue for the purposes of this study.

As depicted in Figure B1.2a, the CC exhibits two distinct peaks on both land and water surfaces. A fixed threshold set to the green line in the figure results in accuracies of 90% and 73% for land and water detection, respectively. The variability in NA over land is smaller compared to that of the CC.

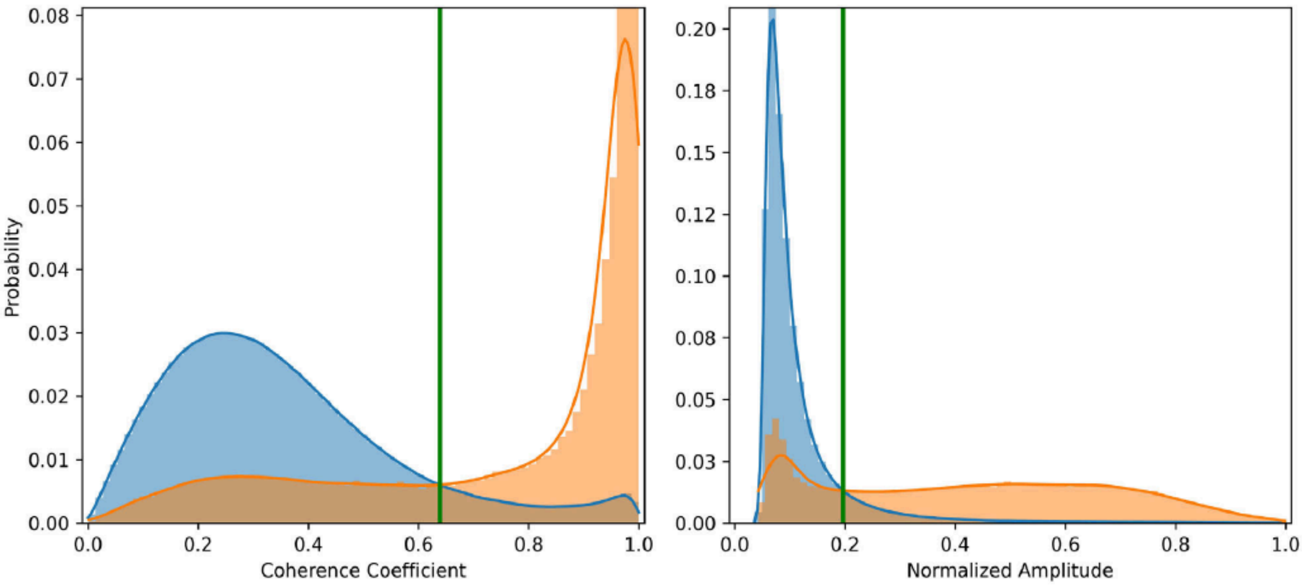


Figure B1.2a: *Distribution of the CC and NA values of the full complex data set, conditioned by presence of surface water (orange) and dry land (blue). Figure from [P24].*

Given that it is likely that the performance of the random forest classifier is affected by the reduced amount of data, we first analyzed the impact of changing the size of the training set. To do so, the random forest is first trained with different proportions of the dataset. The results in Figure B1.2b clearly demonstrate that the larger the size of data utilized for training, the higher the accuracy of the water classification, while the performance of the land classification is less affected. These trends are promising, as many more high-rate complex data will be available in just a few days of HydroGNSS operations. Furthermore, the figure also shows that the inclusion of the NA as an input feature significantly contributes to improving the accuracy of water surface detection (with little impact on the land-surface classification). **Overall, and despite the relatively small dataset, the accuracy of land detection is above 96% and the accuracy of water detection exceeds 90% when using the largest training set and both CC and NA as input features**.

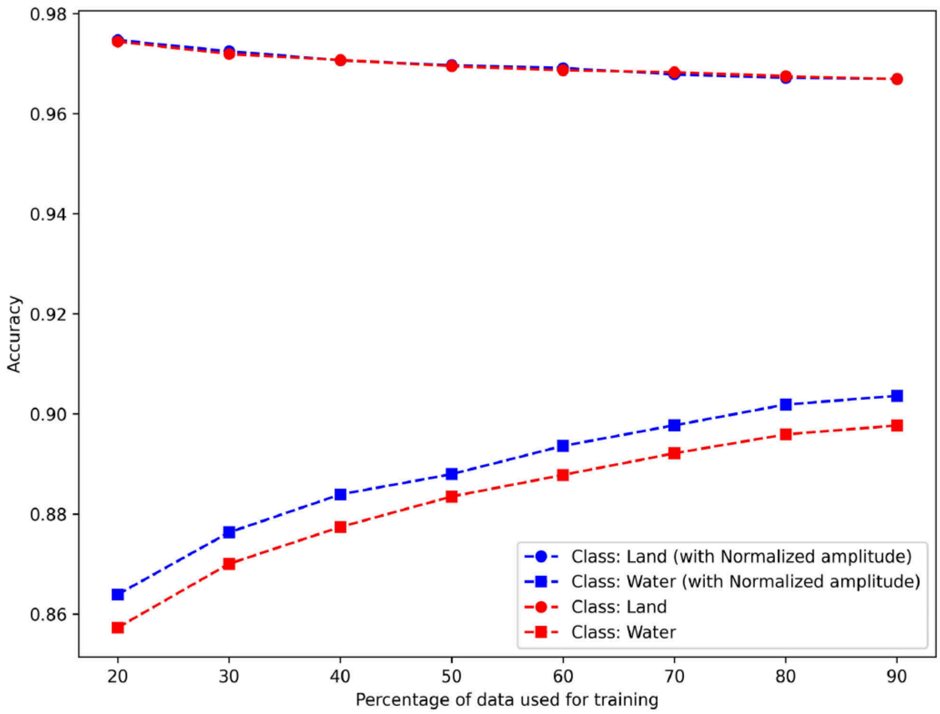


***Figure B1.2b**: Performance of the classifier for high-rate complex signals when different percentages of the total dataset are used as the training set. The classifier has been trained with the CC as only GNSS-R observable (red) and with both CC and NA (blue). Figure from [P24].*

This study also **demonstrates that the algorithms work well when applied to missions with different instrumental characteristics** such as antenna gain and directivity (CYGNSS and TDS1).

B1.3 Validation of polarimetric aspects

A manuscript was prepared and it is currently under review in which the L1B-CC and L2OP-SI/Power algorithms are retrained and applied to dual polarized GNSS-R signals collected from the airborne RONGAWAI mission [P25b]. The main findings of the study are that both reflectivity and PSR are good surface water indicators at LHCP, while at RHCP, the PSR presents stronger sensitivity to the presence of water than the reflectivity. A comparison between the conditioned distribution of PSR at LHCP and RHCP is given in Figure B1.3a.

The performance using dual polarization data was compared to LHCP alone, demonstrating that the incorporation of RHCP can effectively reduce the dependence on auxiliary data. The results from the comparison with flooding products from interferometric synthetic aperture radar and visible infrared imaging radiometer, along with the analysis of feature importance, also indicate that the water probability generated by the classifier is reasonable.

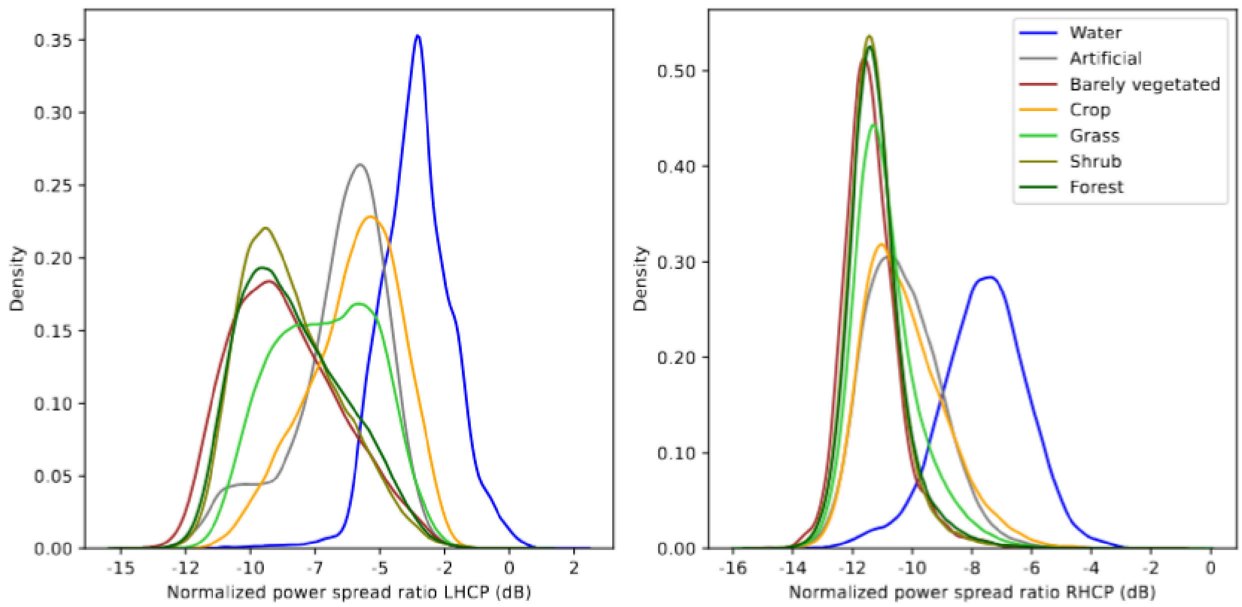


Figure B1.3a*: Distribution of the power spread ratio at LHCP (left panel) and RHCP (right panel) based on RONGOWAI airborne data. Figure from [P25b].*

B.2 Contribution of auxiliary data

One of the concerns of adding auxiliary data in the RF modeling is the actual source of information: are GNSS-R observables driving the output? Or is this mostly predetermined by the auxiliary data?

Given that

1. season, date or time are not given as input variables of the RF algorithm, and
2. among the auxiliary data, only the GNSS observables do change with time

we expect that the time evolution of flooding events is fully driven by the GNSS-R observables. This is illustrated in Figure B12, where a severe flooding event in China is shown.

Another way to quantify the contribution of the GNSS-R observables is through the '*feature_importances_*' fields provided by the Random Forest algorithm. For computational reasons, the beta-version of L2OP-SI uses 5 different RF models, splitting the Globe in 5 big areas, as shown in Figure B11. The 'feature importances' of each of the contributing data, at each of these regions is provided in Table B2.

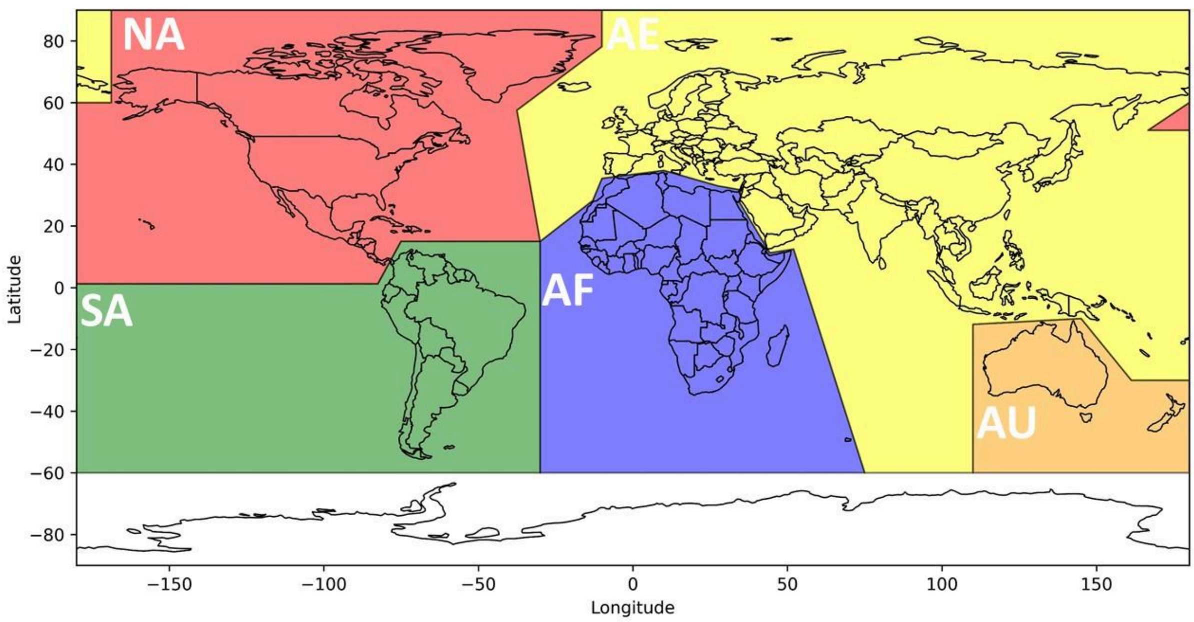


Figure B11: The 5 different areas for beta-version of L2OP-SI.

Table B2: *Contribution of each of the input variables in the L2OP-SI algorithm (beta version). This is given in 5 big regions of the Globe due to computational reasons. Figure B13 shows the aggregated and averaged contributions in a pie chart.*

Region (MinLon, MaxLon, MinLat, MaxLat):	Contribution from different variables:					
	Lon	Lat	GNSS-R PSR	GNSS-R Refl	LandCover	Roughness
NA	0.0414605	0.02203744	0.10655539	0.22437381	0.44777832	0.15779455
SA	0.02196312	0.02541078	0.11741392	0.29391736	0.44153174	0.09976308
AF	0.05268849	0.03461207	0.08643766	0.15620365	0.46350334	0.20655478
AE	0.0640296	0.02467156	0.11550528	0.18621392	0.41119345	0.19838619
AU	0.05134472	0.05935154	0.08284703	0.1654057	0.45319524	0.18785577

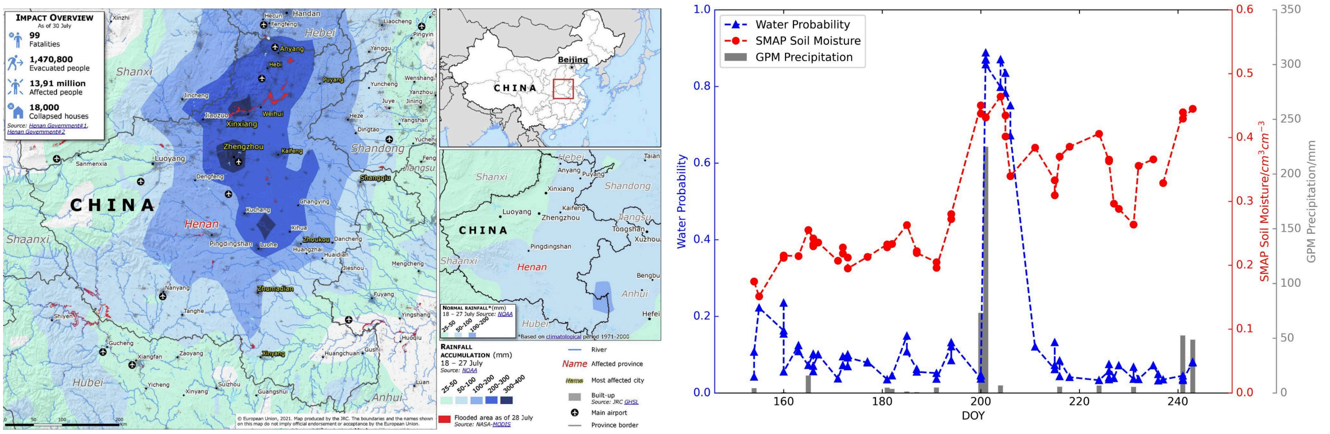


Figure B12: *Right: map of the precipitation regime and flooded areas (in red) in Henan province in Summer 2021 (image by ECHO/European Commission). Left: Water probability computed by the L2OP-SI algorithm on CyGNSS data in Henan province of China in Summer 2021. The plot shows the precipitation (grey), L2OP-SI water probability (blue) and SMAP soil moisture (red) over a small area (~6km x 6 km) in Henan. Severe flooding occurred starting on Day of Year 200 (July 20, 2021) and persisted during ~1 week. The time-dependent water detection in the L2OP-SI algorithm is only driven by GNSS-R observables, the only time-dependent variable used by the algorithm.*

Table B2 shows that the coordinates have a small contribution in the model, the latitude with weights between 0.02 and 0.06 and longitudes between 0.02 and 0.07. However, the land cover and roughness does have a relevant weight. On average across all regions, this weight is larger than 50%, meaning that more than half of the information comes from this auxiliary (external) data set (see Figure B13 for globally averaged values). The reason is that the GNSS-R response is highly dependent on the land cover (forest, types of croplands, deserts…) and roughness, while the presence of water is a modulation on top of this big land cover related signal (we could also add here the topography, urban elements, etc).

Without the land cover and roughness, the algorithm performs relatively good, too – but it performs better adding the land cover and roughness. The contributions of the different variables when excluding land cover and roughness are displayed in Table B3, while the performance of the algorithms –including and excluding land cover and roughness – are given in Table B4. Table B3 shows that in 3 out of 5 regions, the major contribution is from GNSS-R observables, while overall, the average indicates 53% vs 47% contribution of GNSS-R observables vs Coordinates.

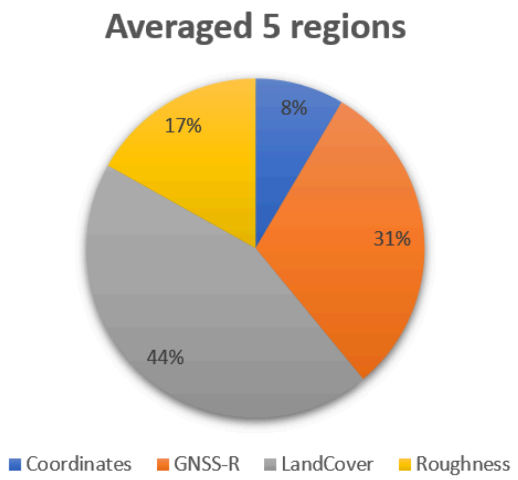


Figure B13: *Pie chart of the contributions ('feature_importance_') of different types of variables into the L2OP-SI Random Forest algorithm. Average values for the 5 different regions detailed in Table B2. The average contribution from the coordinates is 8%, 31% for GNSS-R observables, 44% for the Land Cover and 17% for the roughness.*

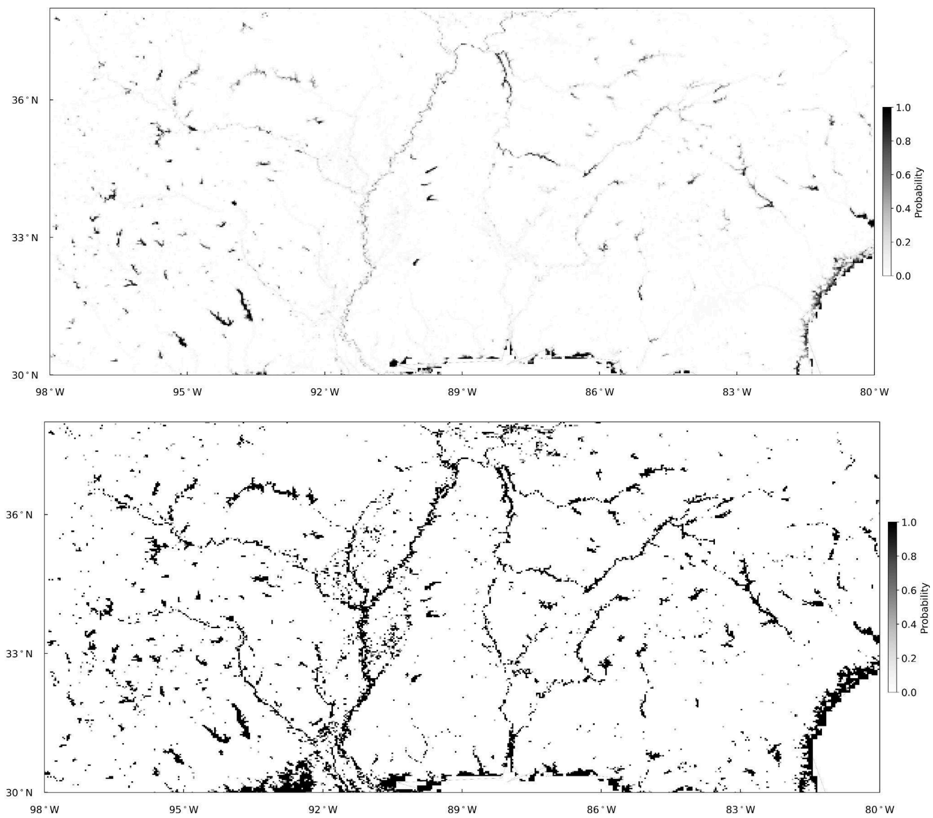
Table B3: *Same as Table B2 but when the algorithm excludes Land Cover and Roughness. The performance of the algorithms worsens a bit, as shown in Table B4.*

Region (MinLon, MaxLon, MinLat, MaxLat):	Contribution from different variables:			
	Lon	Lat	GNSS-R PSR	GNSS-R Refl
-180, 1, -1, 90	0.23430185	0.1765843	0.23171097	0.35740288
-180, 1, -90, 0	0.16073642	0.14119766	0.23037496	0.46769096
0, 51, -1, 90	0.2756838	0.33456243	0.12415505	0.26559873
50, 180, -1, 90	0.28447193	0.18332556	0.18332923	0.34887329
0, 180, -90, 0	0.25079219	0.29443871	0.13112095	0.32364815

Table B4: *Comparison between the overall performance of the L2OP-SI beta processor and that of the algorithm excluding Land Cover and Roughness.*

	L2OP-SI beta	**Excluding Land Cover and Roughness**
Overall accuracy:	0.95	0.895
Accuracy on actual land:	0.95	0.894
Accuracy on actual water:	0.95	0.960

The water probability incoherent from L2OP-SI beta and the model excluding GNSS-R observables were compared, as illustrated in Figure B14. It can be observed that the model excluding GNSS-R observables can provide a rough outline of the water surface, and after incorporating GNSS-R observables, the obtained results are closer to the reference data.



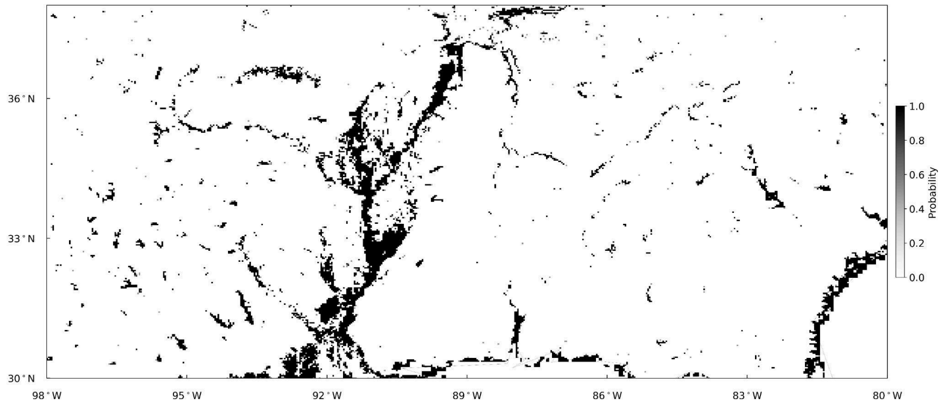


Figure B14: *Top: water probability from reference data,GSW, in September 2021; Middle: water probability from L2OP-SI beta; Bottom: water probability from Model excluding GNSS-R observables.*