STUDY ON HOMOGENIZATION OF SYNTHETIC GNSS-RETRIEVED IWV TIME SERIES AND ITS IMPACT ON TREND ESTIMATES WITH AUTOREGRESSIVE NOISE

Anna Klos[1], Roeland Van Malderen[2], Eric Pottiaux[3], Olivier Bock[4], Janusz Bogusz[1], Barbara Chimani[5],
Michal Elias[6], Marta Gruszczynska[1], Jose Guijarro[7], Selma Zengin Kazanci[8] and Tong Ning[9]

1) Military University of Technology, Warsaw, Poland, 2) Royal Meteorological Institute of Belgium (RMI), Brussels, Belgium, 3) Royal Observatory of Belgium (ROB), Brussels, Belgium,
4) IGN LAREG, University Paris Diderot, Sorbonne Paris, France, 5) Central Institute for Meteorology and Geodynamics, Austria, 6) Research Institute of Geodesy, Topography and Cartography, Czech Republic,
7) AEMET (Spanish Meteorological Agency), Spain, 8) Karadeniz Technical University, Turkey, 9) Lantmäteriet, Sweden

European Geosciences Union General Assembly 2017
Vienna | Austria | 23–28 April 2017

contact e-mails: anna.klos@wat.edu.pl, roeland@meteo.be, Eric.Pottiaux@oma.be

ABSTRACT

A synthetic benchmark dataset of Integrated Water Vapour (IWV) was created within the activity of "Data homogenisation" of sub-working group WG3 of COST ES1206 Action. The benchmark dataset was created basing on the analysis of IWV differences retrieved by Global Positioning System (GPS) International GNSS Service (IGS) stations using European Centre for Medium-Range Weather Forecats (ECMWF) reanalysis data (ERA-Interim). Having analysed a set of 120 series of IWV differences (ERAI-GPS) derived for IGS stations, we delivered parameters of a number of gaps and breaks for every certain station. Moreover, we estimated values of trends, significant seasonalities and character of residuals when deterministic model was removed. We tested five different noise models and found that a combination of white and autoregressive processes of first order describes the stochastic part with a good accuracy. Basing on this analysis, we performed Monte Carlo simulations of 25 years long data with two different types of noise: white as well as combination of white and autoregressive processes. We also added few strictly defined offsets, creating three variants of synthetic dataset: **easy**, **less complicated** and **fully complicated**. The synthetic dataset we present was used as a benchmark to test various statistical tools in terms of homogenisation task. In this research, we assess the impact of the noise model, trend and gaps on the performance of statistical methods to detect simulated change points.

IWV DATA & HOMOGENISATION

IGS „**repro1**" troposphere products screened and converted to **Integrated Water Vapour (IWV)** by O. Bock (https://doi.org/10.14768/06337394-73a9-407c-9997-0e380dac5590). A set of 120 stations with daily observations from a period of 1995-2010 (Figure 1). We focused on the differences: ERAI-GPS, which may reveal artificial breaks (Figure 2).

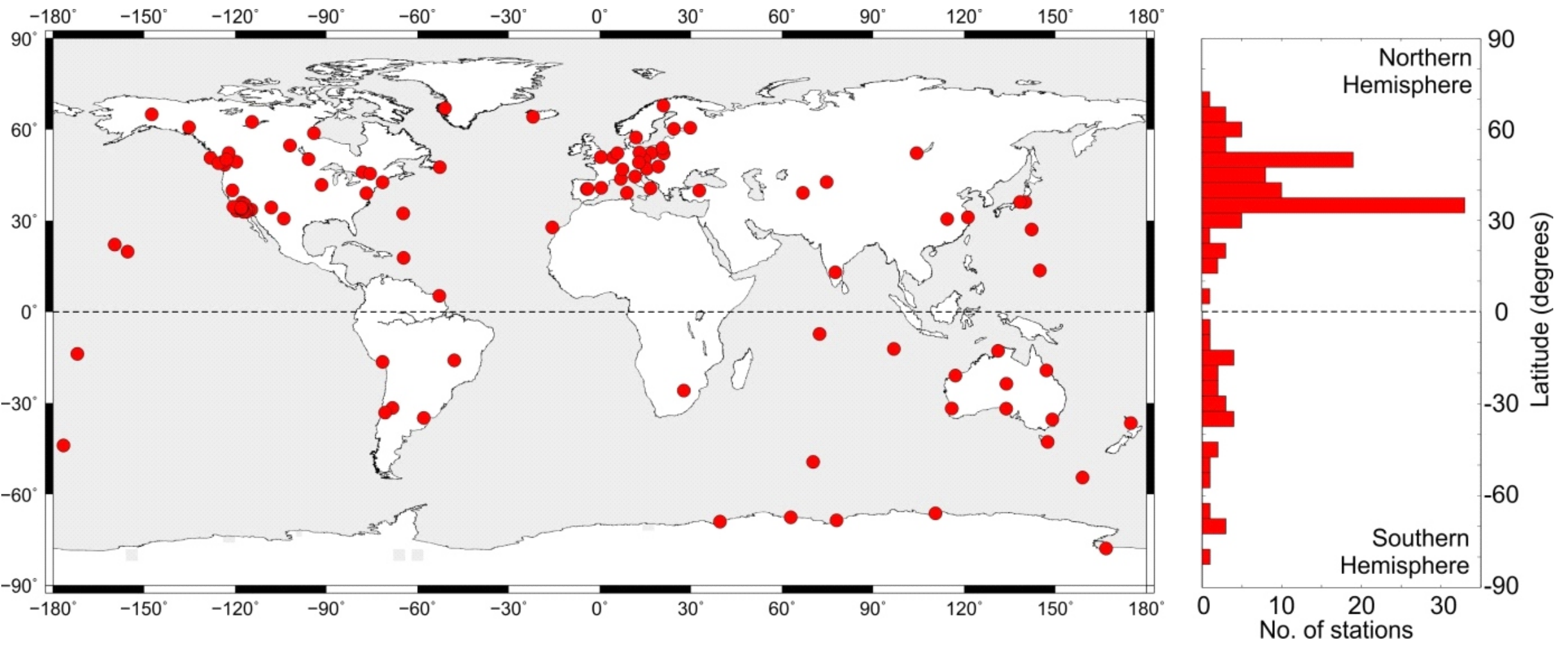


Figure 1. *A set of 120 permanent IGS stations processed within „repro1" campaign.*

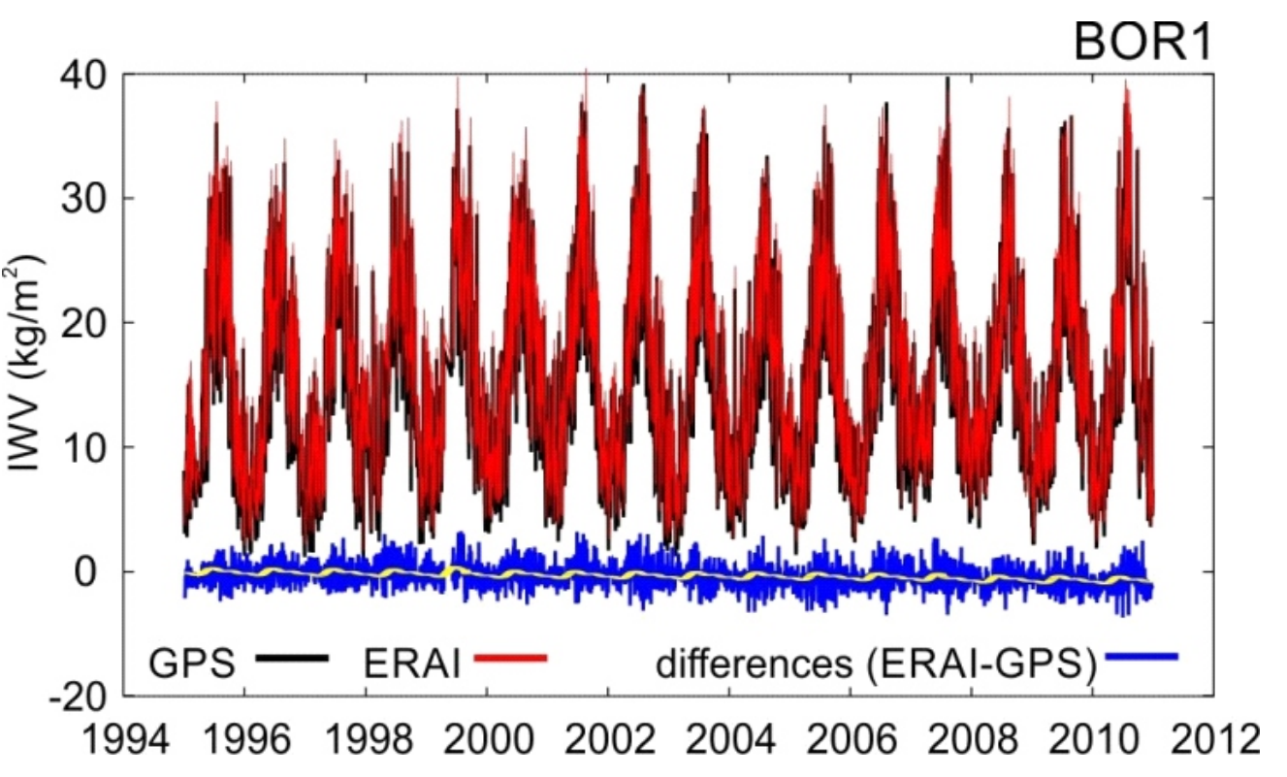


Figure 2. *IWV retrievals from GPS and ERAI model estimated for BOR1 (Poland) station.*

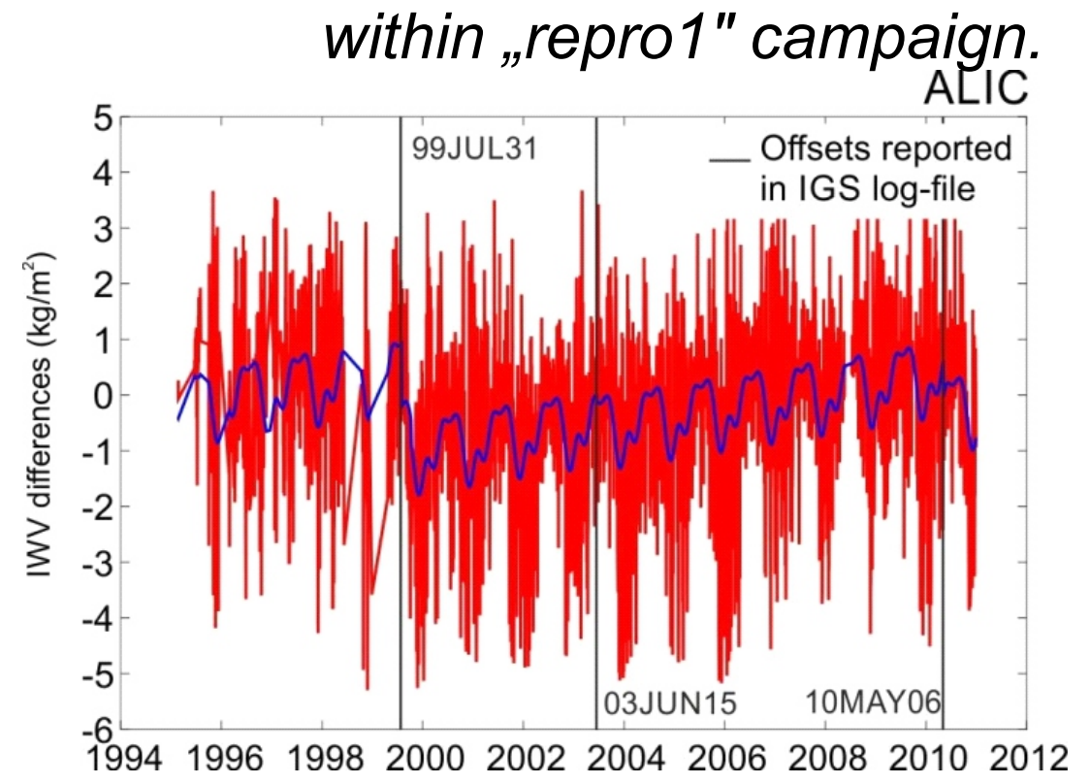


Figure 3. *Offsets reported for ALIC (Australia) permanent station. Blue line indicates the least-squares model fitted into series.*

Manual homogenisation (Figure 3)
Before we delivered the parameters of deterministic and stochastic parts to create a synthetic benchmark, we performed a manual homogenisation, as undetected breaks might influence further estimates.

1. 1029 epochs which were reported in the IGS log files (ftp://igs.ign.fr/pub/igs/igscb/station/log/ and ftp://igs.ign.fr/pub/igs/igscb/station/oldlog/, changes in R=receiver, A=antenna, D=radome).
2. We averaged the epochs from log files from 1029 to 970 epochs: 164 offsets validated manually.
3. 57 offsets were additionally reported during manual checking. Finally: **221 epochs**.

ANALYSIS OF TIME SERIES OF IWV DIFFERENCES

IWV differences are characterized by trends, amplitudes, seasonal signals and noise. We found that a combination of autoregressive process of first order and white noise is the optimal process for IWV differences (Figure 4). We delivered the noise parameters in a form of: fraction of AR(1), coefficient of AR(1) and amplitudes of WN and AR(1) noise (Figures 5 & 6). All parameters which characterized individual stations were employed to create a synthetic benchmark.
Computations were performed with Maximum Likelihood Estimation (MLE) ans a model of:

$$IWV(t_i)=a+b(t_i)+c\sin(2\pi t_i)+d\cos(2\pi t_i)+e\sin(4\pi t_i)+f\cos(4\pi t_i)+g\sin(6\pi t_i)+h\cos(6\pi t_i)+k\sin(8\pi t_i)+l\cos(8\pi t_i)+\sum_{j=1}^{n}m_jH+\varepsilon_{IWV_i}$$

Figure 4. *Combination of autoregressive process of first order and white noise* ***[AR(1)+WN]*** *characterized by: AR: a coefficient, fraction and amplitude, WN: amplitude was found* ***to be optimal****.*

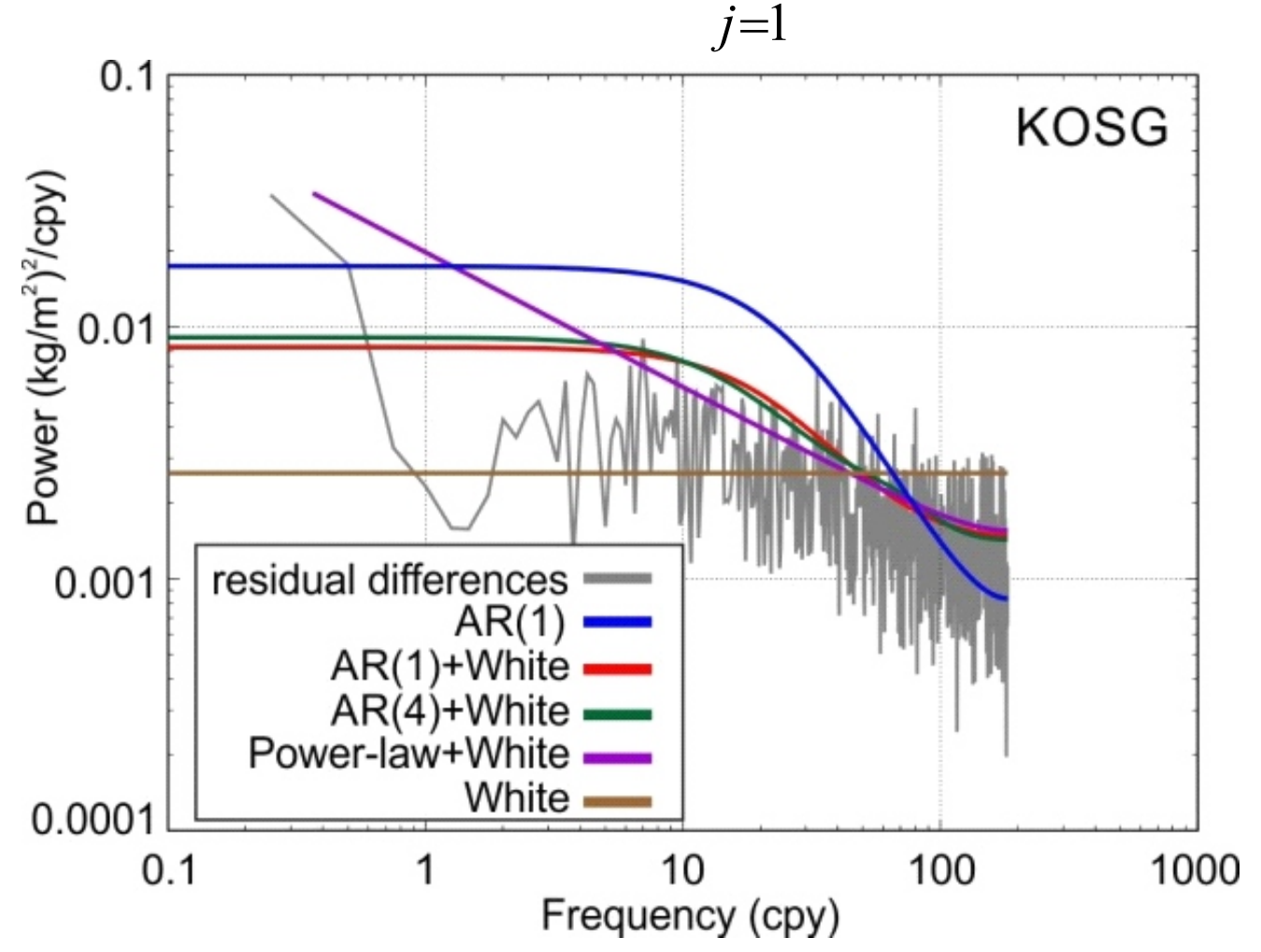
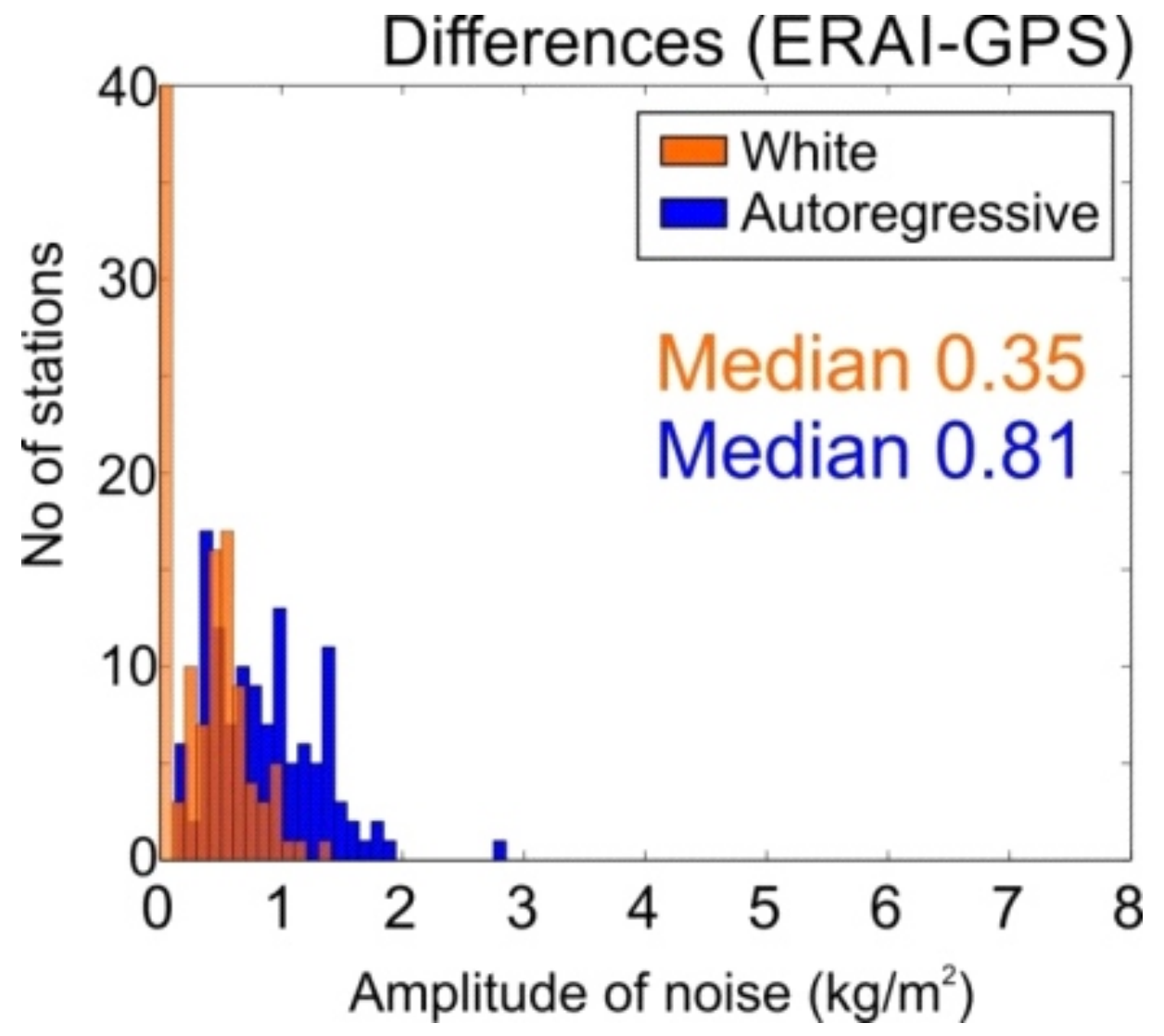


Figure 5. *The amplitudes of autoregressive process of first order and white noise estimated for IWV differences for a set of 120 IGS permanent stations.*



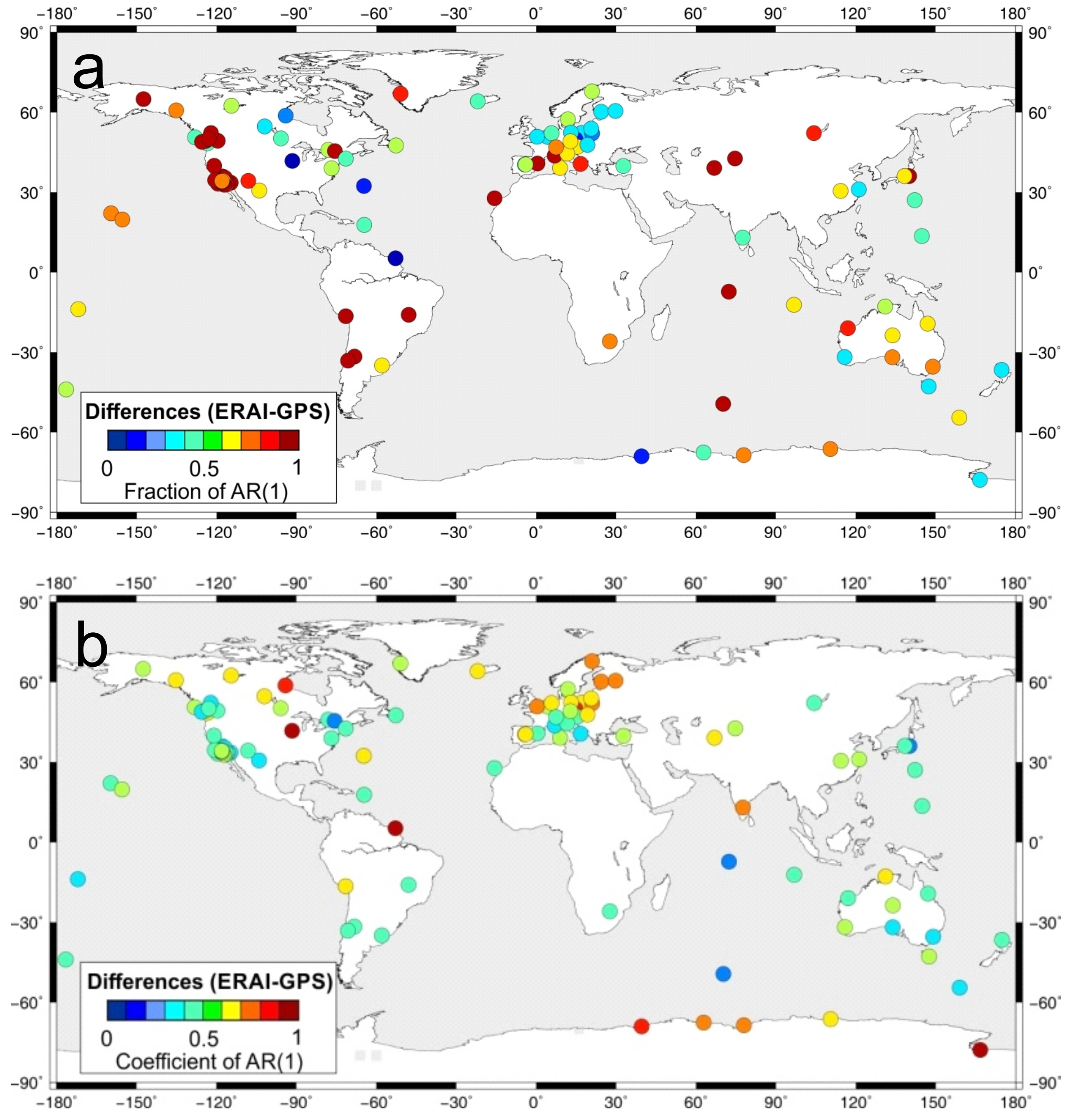


Figure 6. *Fraction (a) and coeffients (b) of AR(1)+WN noise model.*

GENERATION OF SYNTHETIC BENCHMARK

We created 3 variants of synthetic time series separately for each station (120 series for each variant):
1. **EASY** dataset: seasonal signals (annual, semi-annual, ter- and quater-annual, if found for a particular station) + offsets + white noise (WN),
2. **LESS COMPLICATED** dataset: same as 1. + autoregressive process of the first order (noise model = AR(1)+WN),
3. **FULLY COMPLICATED** dataset: same as 2. + trend + gaps (up to 20% of missing data).

Generation of the benchmark dataset:
- mathematical model of data taken directly from real differences: trend, seasonal signals, noise,
- offsets simulated randomly,
- number of offsets and exact epochs **were blinded for potential users**.

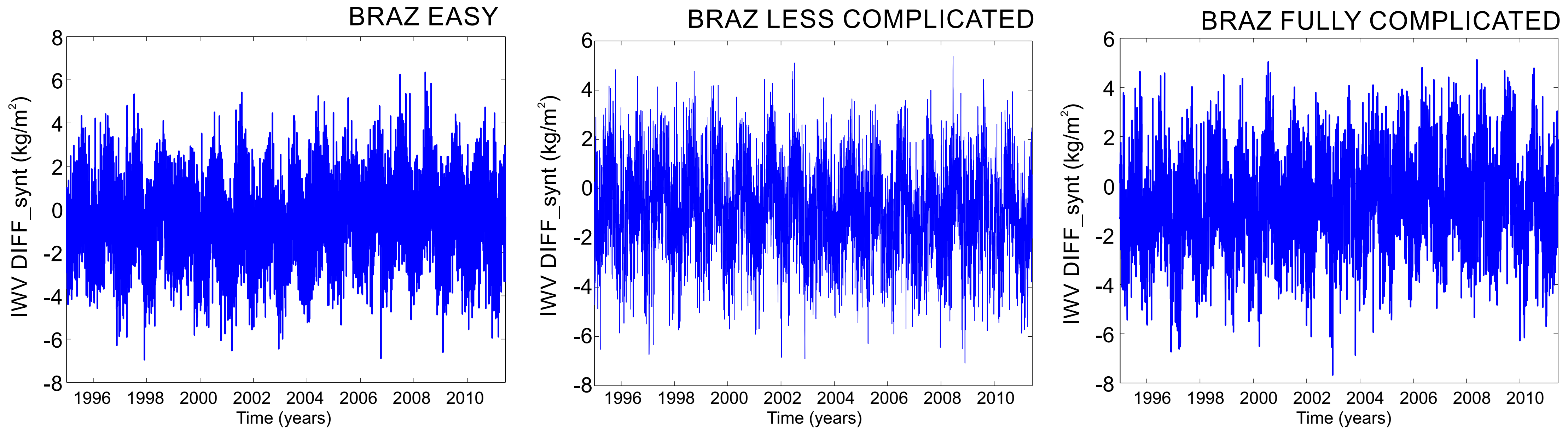


Figure 7. *Easy, less and fully complicated datasets created for parameters delivered for station BRAZ (Brazil).*

COMPARISON OF THE STATISTICAL TOOLS

We tested various statistical approaches to detect the simulated offsets. On the basis of the number of True Positives, False Positives and False Negatives, we assessed how good the statistical methods perform under varying noise and time series characteristics (Figures 8 & 9).

	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7
Symbol	● ⬬	▲	+	×	◆	▼	-
Operator	M. Elias	R. Van Malderen	R. Van Malderen	J. Guijarro	T. Ning	S. Zengin	B.Chimani
Method / SW	2-sample t-test	2 of 3	PMW	CLIMATOL	PMTred	Pettitt	HOMOP
Daily/Monthly	D+M	D+M	D+M	D+M	D+M	D	X
Easy/Less/Full	E+L+F	E+L+F	E+L+F	L+F	E+L+F	E+L+F	E+F

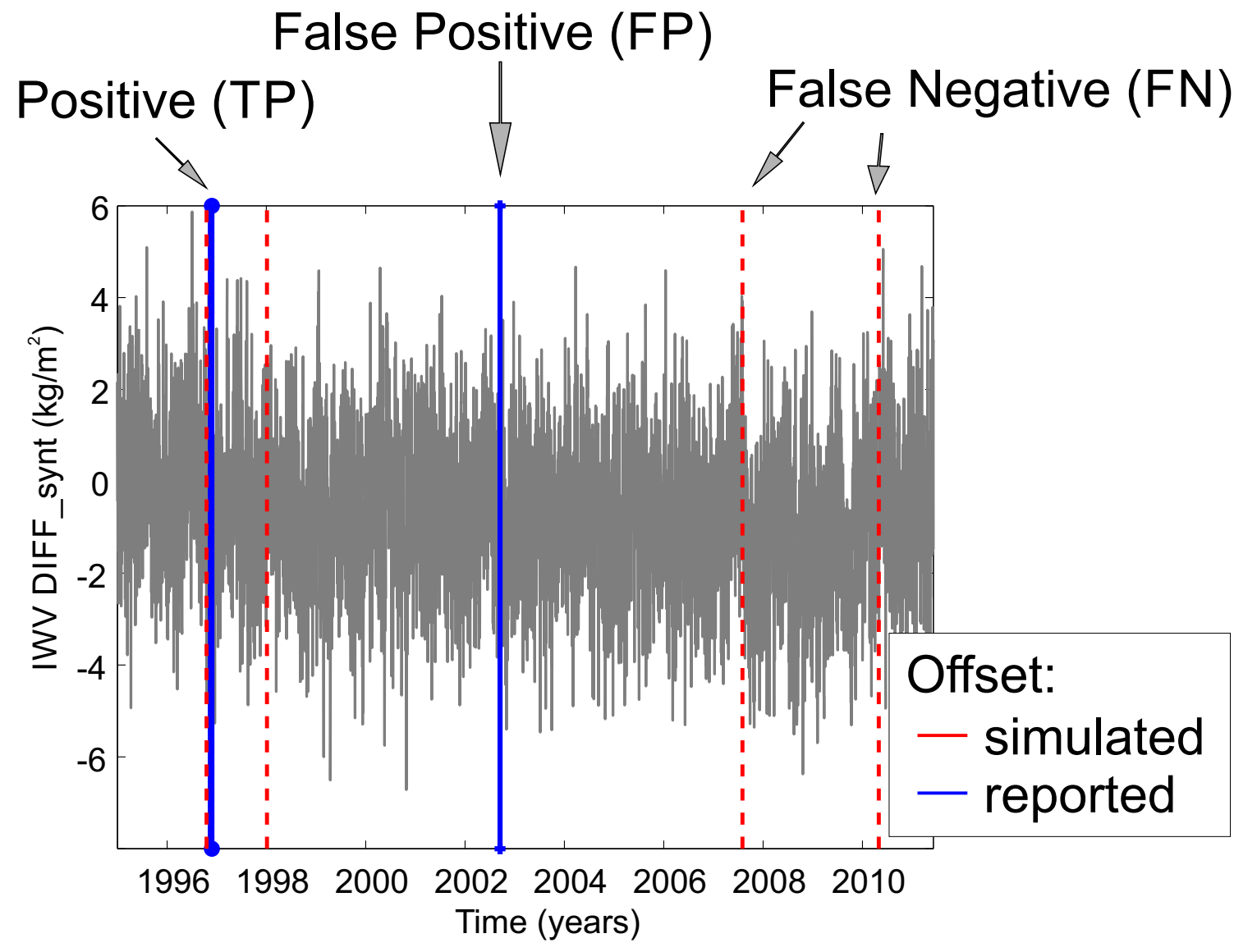
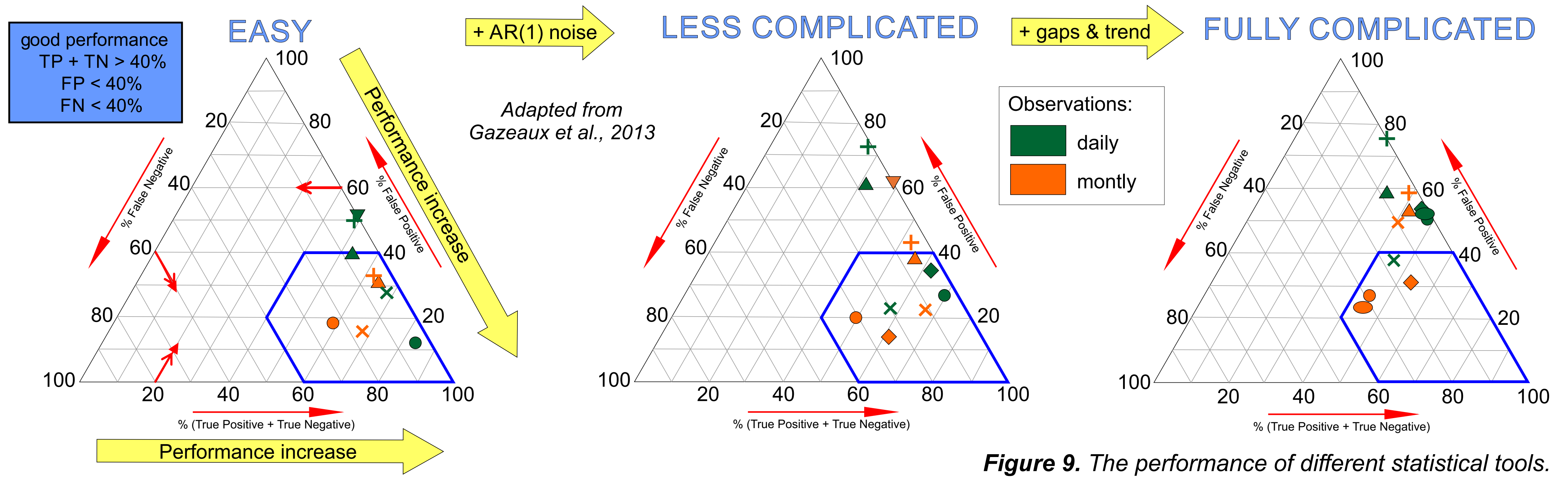


Figure 8. *The idea of True Positives (TP), False Positives (FP) and False Negatives (FN). True Negatives (NS) (not shown here) mean no break was present and no break was detected.*



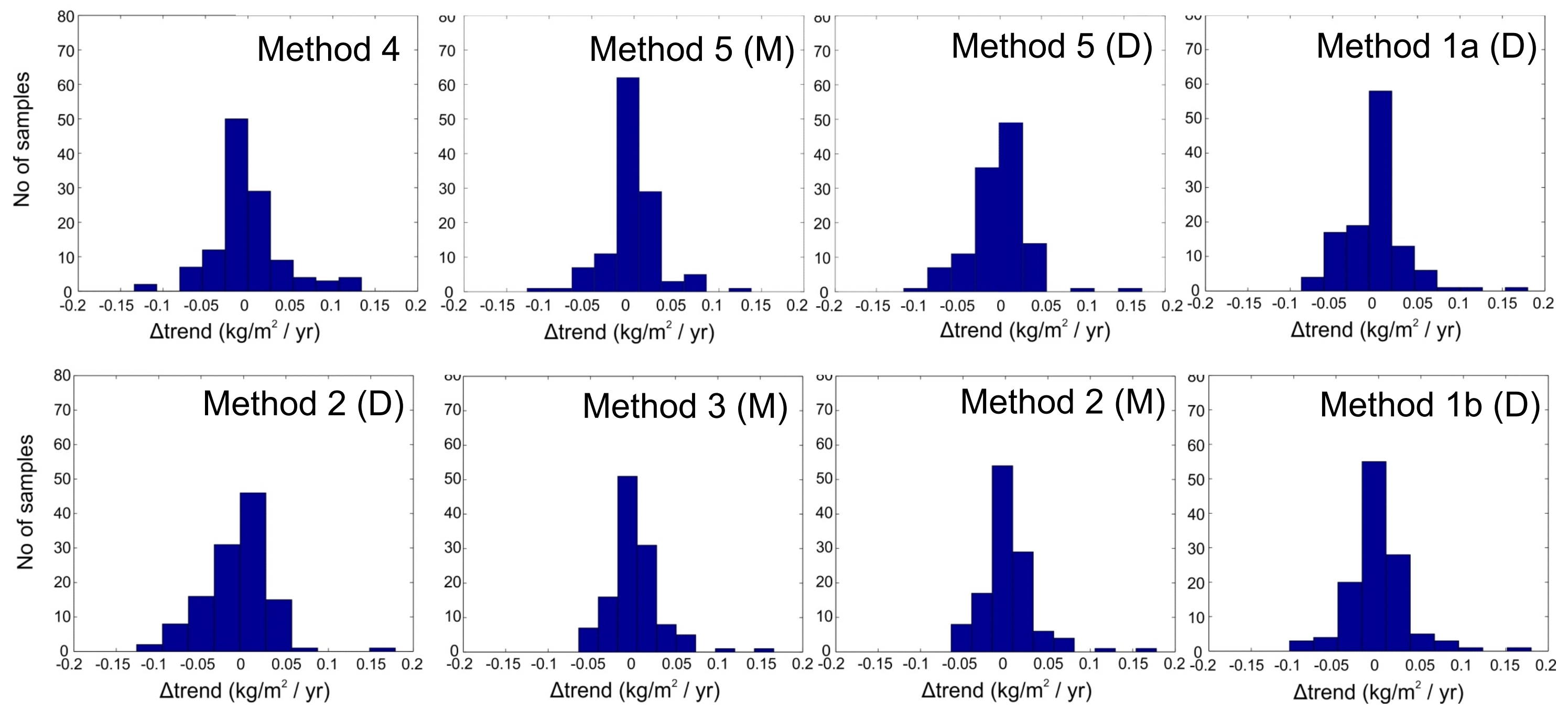
good performance
TP + TN > 40%
FP < 40%
FN < 40%

Adapted from Gazeaux et al., 2013

Figure 9. *The performance of different statistical tools.*

DIFFERENCES IN TREND (FULLY COMPLICATED VARIANT)

We estimated the changes in trend between simulated series and the one delivered after epochs of offsets were reported by different tools (Figure 10).



Number of reported offsets in fully complicated dataset:

Method 1a (D): 170	Method 3 (M): 260
Method 1b (D): 185	Method 4 (D): 146
Method 2 (D): 386	Method 5 (D): 264
Method 2 (M): 238	Method 5 (M): 128

Figure 10. *The differences in values of trend between synthetic benchmark and the estimated trend when different statistical tools were applied.*

We are looking forward to getting feedback/input/participation from your community!

PLANS FOR THE FUTURE

CONTACT US!

The epochs of offsets were revealed. Now, there is a time for participants to fine-tune their tools. Then, a next generation of a fully complicated synthetic dataset will be generated and a blind homogenisation re-run again.

ACKNOWLEDGMENTS
Anna Klos was supported by COST Action ES1206 GNSS4SWEC (gnss4swec.knmi.nl) during her stay at the Royal Observatory of Belgium (ROB) and the Royal Meteorological Institute of Belgium (RMI).
Janusz Bogusz is supported by the Polish National Science Centre, grant. No. UMO-2016/21/B/ST10/02353.