



**Politecnico
di Torino**

Politecnico di Torino

Corso di Laurea Magistrale in Ingegneria Matematica

Modelling, Generation and Mitigation of Striping Noise in the ESA MetOp-SG Microwave Imager

Satellite Destriping Scientific Challenge Preparation

Relatore:
Enrico Magli

Candidata:
Monica Beghini

17 Luglio 2026

Abstract

Over the past decades, Earth Observation satellites have become essential across many fields, in particular weather forecasting and climate monitoring. As part of this continuous advancement, the MetOp Second Generation satellites, Europe’s next-generation polar-orbiting spacecraft developed by ESA and EUMETSAT, will provide observations of substantially improved quality for more than two decades. Among the instruments they carry, the Microwave Imager (MWI) is affected by striping patterns: a low-frequency flicker component in the receiver electronics produces long-range temporal correlations, which manifest as local excursions in the signal and appear in the imagery as a coherent directional pattern of stripes. As the brightness temperatures measured by the instrument are directly assimilated into numerical weather prediction models, this spatial artefact risks being misinterpreted as a genuine geophysical feature, degrading downstream applications and motivating the need for dedicated destriping algorithms.

As this issue is generic to microwave instruments, the goal of this thesis is to prepare the dataset and the performance baseline for a scientific challenge, an open competition through which the research community can develop novel destriping methods.

The work proceeds in three stages. First, the MWI instrument noise is characterised from ground-test data, yielding a statistical model that describes it as a Gaussian, two-component process – a white thermal floor and a long-memory flicker component – with a reproducible power spectral density. Second, this characterisation is used to build two synthetic noise generators, a parametric and an empirical one, both exploiting the stability of the spectral structure to reproduce the striping. Finally, three destriping algorithms are implemented, tested and compared. A hybrid denoising framework combining two of the selected methods is proposed and shown to reduce striping artefacts more effectively than either method alone while preserving the underlying weather signal.

The synthetic noise is then injected into simulated MWI Earth orbits to produce the challenge dataset, and the proposed hybrid method is used to define its baseline against which participant submissions should be evaluated.

Table of Contents

1	Introduction	1
1.1	The Instrument	2
1.2	Problem Statement	3
1.3	Objectives	5
1.4	Outline	6
2	State of the Art	7
2.1	Denoising and Destriping	7
2.2	Striping in Satellite Instruments	8
2.3	Destriping Methods	9
3	Instrument Noise Modelling	12
3.1	Time Series Analysis of Instrument Noise	12
3.1.1	Ground-Testing Setup and Data	12
3.1.2	First and Second Order Statistics	13
3.1.3	ACF, Allan Deviation and PSD	17
3.1.4	Noise Model	22
3.2	Spatial Manifestation of Noise: Striping	24
3.2.1	Striping Metrics and Quantification	24
3.2.2	Operational Calibration Process	25
3.2.3	Statistical Analysis of Stripe Occurrence	29
4	Synthetic Noise Generation	36
4.1	Thermal+1/f model	36
4.1.1	Cut-off frequency implementation	37
4.2	LUT model	38
4.2.1	Window Length Selection	38
4.3	Models Comparison	40

5	Noise Mitigation	43
5.1	Wiener Filter	44
5.2	PCA/EEMD	46
5.3	BM3D	50
5.4	Hybrid Method: PCA/EEMD+BM3D	53
6	Challenge Preparation and Results	54
6.1	Challenge Dataset	54
6.1.1	MWI L1B Test Data	54
6.1.2	Dataset Generation	56
6.2	Methods Evaluation	59
6.2.1	Metrics	59
6.2.2	Per-Method Analysis	60
6.3	Challenge Baseline	69
6.3.1	Scene-Adaptive Score	69
6.3.2	Baseline Evaluation	71
6.4	Train–Test Split	75
6.5	A Structure-Aware Hybrid	75
7	Conclusions	78
7.1	Summary of Contributions	78
7.2	Limitations and Future Work	79
	Bibliography	81
	Acknowledgements	85

Chapter 1

Introduction

Since the mid-twentieth century, the science of Earth Observation (EO) satellites has developed from low-resolution images used mainly for military reasons to the daily acquisition of data which allows the monitoring, measuring and understanding of Earth-system processes. The impact of EO spans across several fields, from geology to map creation, and amongst them clearly meteorology [1].

Although the possibility to forecast the weather by solving a system of nonlinear partial differential equations was first proposed by Vilhelm Bjerknes in 1904, several decades passed before Numerical Weather Prediction (NWP) was actually feasible. The two main limiting constraints were the amount of computer power, or the lack thereof before the first computer, and the amount of data available to initialize the mathematical models [2].

In the same way the former was solved by technological progress, the latter increased significantly through remote sensing: in the present day, the vast majority of observations used in NWP are indeed obtained from satellites.

Among the instruments of meteorological satellites, microwave imagers play an important role: by sensing microwave radiation coming from the Earth's surface and the atmosphere, they provide information on temperature and water vapour, on the cloud microphysics and water content, and on surface properties such as sea ice, snow cover and precipitation over land and ocean. As they allow the monitoring of climate variables such as clouds, humidity and precipitation, continued access to high-quality microwave imagery is central to maintaining the forecast accuracy and climate monitoring capacity that modern operational services depend on [3].

1.1 The Instrument

MetOp-SG (Meteorological Operational Satellites - Second Generation) is Europe's next generation of polar-orbiting satellites. Following the first generation, MetOp-SG satellites will provide observations with substantially better coverage and accuracy, with data of crucial importance both for forecasts and environmental monitoring [4]. The programme, developed in collaboration between the European Space Agency (ESA) and EUMETSAT, includes two types of satellites working as a pair: MetOp-SG A and MetOp-SG B, each type carrying a different but complementary instrument package, for a total of 10 sensor instruments in all [5, 6]. Three successive pairs of MetOp-SG A and MetOp-SG B spacecraft will be launched approximately every seven years, enabling full operational coverage for more than two decades [7]. Currently, only MetOp-SG A1 is in orbit, launched on 13 August 2025; the next satellite will be MetOp-SG B1, scheduled for launch by the end of 2026.

The object of this thesis is the Microwave Imager (MWI), a conically scanning radiometer developed by ESA and carried by the B-type satellites; the MWI measures thermal radiance emitted by the Earth at high spatial resolution in the microwave region of the spectrum. Unlike most MetOp-SG instruments, the MWI has no direct predecessor on the first-generation MetOp satellites, introducing a microwave imaging capability the earlier meteorological programme did not carry. As already anticipated, the objective of such an instrument is to support NWP by providing cloud and precipitation products, all weather surface imagery and to obtain sea surface winds and total column water vapour above oceans; as for very short-range forecasting at regional scales, its goal is to reach key requirements on cloud liquid water and ice estimates, and precipitation estimates.

The MWI has the following instrument specifics [8]:

- *Spectral.* The frequency coverage is from 18GHz up to 183GHz, for a total of 18 frequencies over 26 channels; the first 8 frequencies are measured with both vertical (V) and horizontal (H) polarizations, while the following 10 with V polarization only. The focus of this work will be the single MWI-13 V channel, the 21st channel with central frequency at 165.5 ± 0.75 GHz; together with the channels up to the 26th, they are less sensitive to the surface and indeed their aim, thanks to cloud slicing, is to provide information on water vapour profiles and snowfall.
- *Scanning.* The conical scanning is performed by means of a rotating offset parabolic reflector antenna and feed-horn cluster rotating together at 45 rpm, i.e. one revolution every 1.33 s. As illustrated in Figure 1.1, during each revolution the instrument observes two calibration targets and a portion of the

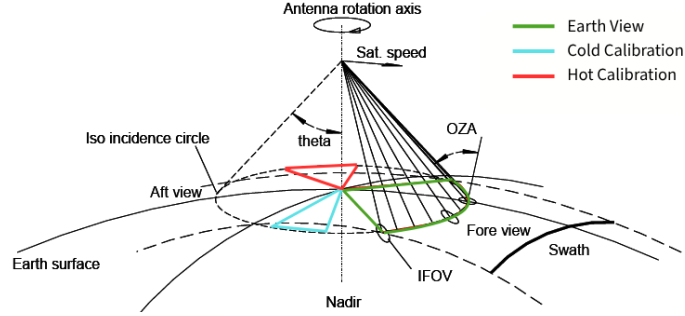


Figure 1.1: MWI viewing geometry. Source: adapted from [8].

Earth's surface. The two references are deep space, the cold target at about 3 K, and the On-Board Calibration Target (OBCT), the hot target at about 280 K. The Earth view portion spans a swath of roughly 1700 km and, at a sampling time of 0.394 ms, yields 1394 samples (a value still marked to-be-confirmed, see Chapter 6). A full orbital revolution takes about 101 minutes.

1.2 Problem Statement

Extensive testing of all instruments and satellites is performed on ground before launch. As for the MWI instrument, the radiometric performance and calibration tests are of crucial importance: using On-Ground Calibration Targets (OGCT), temperature controlled microwave blackbodies that allow radiofrequency stimulation, they ensure the requirements for the radiometric performance are met. Specifically, two kinds of test target are needed [9]:

- *Ambient targets.* At fixed temperature, they represent a known reference in ambient tests, to characterise the radiometric gain and sensitivity of the instrument.
- *Thermal-vacuum targets.* They simulate the *Cold Sky* at fixed temperature and the *Earth* at variable temperature, to characterise the radiometric sensitivity and bias.

During the MWI Proto-Flight Model (PFM) measurements, all channels were compliant with margin except one: the MWI-13 V channel registered an unsatisfactory performance, with *striping* patterns appearing in the imagery. This thesis therefore focuses on the single MWI-13 V channel; all noise characterisation and mitigation in the following chapters refers to this channel, while MWI continues to denote the

instrument as a whole.

As will be described in Chapter 2, striping is a common problem in imaging and, although it arises from various causes and its appearance varies accordingly, it always manifests as a directional pattern of brighter and darker lines.

In the case of satellite imaging, the directionality is given by the scanning motion, and in particular two directions can be distinguished:

- *Along Track*. It follows the satellite’s motion along its orbit; adjacent along-track samples are separated by the time required to complete one full antenna rotation (1.33 s for MWI).
- *Cross Track*. It is spanned within a single rotation of the antenna and the sampling time determines the separation between two adjacent samples (0.394 ms for MWI); in the image produced by the acquisition, it is perpendicular to the along-track direction.

Origin of Striping

The physical origin of striping lies in the receiver electronics: alongside the unavoidable *thermal* noise, caused by the random motion of agitated charge carriers, there is the *flicker* noise (also $1/f$ noise), which is most pronounced in instruments sensitive to surface phenomena related to charge trapping and releasing [10]. Such noise is defined spectrally with a $1/f$ slope, dominating in the low-frequency regime. In the time domain, this concentration of power at lower frequencies produces long-range temporal correlations, causing excursions in the signal anywhere from a few milliseconds to tens of seconds. While cross-track samples are only milliseconds apart and therefore share essentially the same instantaneous noise level, along-track samples belonging to successive rotations are seconds apart and can have wandered to very different noise levels. When the time series is folded into the two-dimensional swath image, this phenomenon manifests as a coherent offset shared by the samples of a single scanline; the offset varies from one scanline to the next, producing a striping pattern that develops in the along-track direction.

Radiometric Calibration

The two reference targets observed in each scan – deep space and the OBCT introduced in Section 1.1 – serve the purpose of converting the raw detector counts into calibrated brightness temperatures (BT) through a two-point calibration procedure. In a full 360° antenna rotation, the Earth scene accounts for 41.22% of the duty cycle, while 20.45% is devoted to the two references. Their measured counts and known temperatures define a linear count-to-temperature relation whose slope is applied uniformly to all cross-track samples of the Earth view that follows, until

the next calibration event.

The calibration procedure interacts with the striping problem in two opposing ways. On one hand, the calibration slope estimate, i.e. the inverse of the receiver gain (hereafter simply gain), is computed from noisy reference counts, so any noise present during the reference observations propagates directly into the gain and from there into every calibrated BT value of that scanline. On the other hand, by re-estimating the gain at every scanline, calibration resets the slow drift periodically, reducing the along-track fluctuation it would otherwise accumulate. Overall, as will be discussed in Section 3.2.2, calibration mitigates but cannot resolve the striping problem. The fundamental limitation is statistical: thermal noise is uncorrelated from sample to sample, so its effect on the gain estimate can be reduced by averaging more reference counts; the wandering produced by flicker noise, however, is only partially reduced by averaging since it is caused by power concentrated at lower frequencies. Furthermore, even a perfectly estimated gain cannot correct an excursion that occurred during the Earth view acquisition itself.

Striping is not merely a visual artefact: the brightness temperatures produced by microwave imagers are directly assimilated by NWP models and used to retrieve the geophysical quantities required for weather forecasting and climate monitoring. Because striping introduces coherent spatial structures, it risks being misinterpreted as genuine features in these downstream applications, corrupting retrievals and acting as a persistent source of degradation. Characterising and mitigating striping noise at the calibration or post-processing stage is therefore essential to preserve the scientific value of the observations.

1.3 Objectives

Striping is an intrinsic property of instruments such as the MWI, and for this reason technological improvements aimed at reducing its severity at the instrument level are actively being pursued. However, even an instrument that meets its requirements at launch may degrade in orbit: noise levels can increase substantially over the course of a mission, as has been documented for several operational microwave sounders [10]. The problem is indeed not unique to MWI – analogous striping artefacts have been observed across a range of similar instruments, suggesting a common underlying vulnerability of this class of sensors.

The combination of these factors motivates the need for robust and general noise mitigation strategies. One effective approach to accelerating progress on such complex problems is a scientific challenge: a structured open competition in which the broader research community is invited to develop solutions against a common

dataset and evaluation protocol, maximising the likelihood of innovative approaches being discovered. This thesis constitutes the preparatory work for such a challenge and is organised around two main objectives:

1. *Dataset preparation.* Statistical characterisation of the MWI instrument noise from ground-test data and development of a validated synthetic noise generator for the preparation of the challenge dataset, obtained by injecting the synthetic noise onto clean real or simulated Earth observations.
2. *Baseline definition.* Implementation and evaluation of destriping methods to establish a performance baseline against which challenge submissions can be assessed.

1.4 Outline

The thesis is organized as follows.

Chapter 2 provides a review of the state of the art in destriping methods, which will allow the reader to better understand the striping problem and, as a consequence, the approach taken to both objectives.

Chapter 3 and 4 are dedicated to the first objective. Chapter 3 presents an extensive statistical analysis of the MWI-13 V ground-test data: it first characterises the noise in terms of its first- and second-order statistics, temporal correlation structure and spectral composition, from which a noise model is derived; it then examines the spatial manifestation of the noise as striping, describing the operational calibration process and providing a statistical analysis of stripe occurrence in the calibration and Earth view windows. Chapter 4 builds on this analysis to develop, validate and compare two approaches to synthetic noise generation – a parametric and an empirical one.

Chapter 5 and 6 address the second objective. Chapter 5 describes and implements a set of destriping techniques drawn from different domains, together with a hybrid combination of two of them. Chapter 6 builds the challenge dataset from simulated MWI Level-1B orbits, defines the evaluation metrics and a scene-adaptive score, and uses them to evaluate the methods and fix a baseline; it then analyses the baseline’s residual failure and proposes a candidate refinement that detects it from the noisy data alone.

Finally, Chapter 7 summarises the contributions, discusses the limitations encountered, and outlines the next steps towards the final challenge.

Chapter 2

State of the Art

The mitigation of MWI noise consists in removing the directional, structured artefact known as *striping*, and it is therefore a particular instance of the more general problem of recovering a clean signal from noisy observations.

This chapter first defines the general denoising problem, then narrows it to destriping and, in particular, to satellite imagery. After an overview of the four families of destriping methods, three are selected, based on the instrument’s constraints and a set of design choices.

2.1 Denoising and Destriping

The problem of denoising consists in the recovery of an unknown clean signal X from an observation Y corrupted by noise N . The most widely used model – and the one adopted throughout this work – is additive:

$$Y = X + N. \tag{2.1}$$

The solution consists in an estimate $\hat{X}$ of X from Y alone; such a problem is ill-posed and a unique solution cannot be obtained from the noisy observation without introducing prior knowledge about the signal, the noise, or both [11].

The nature of the prior is therefore a distinguishing feature of a denoising algorithm: *noise-agnostic* methods make assumptions only about the underlying signal – most commonly smoothness and local continuity; *noise-aware* methods exploit instead an explicit characterisation of N to estimate the solution.

In the case of MWI, a noise model is derived in Chapter 3, so a noise-aware method is expected to bring an advantage over a noise-agnostic one; this is also a point of agreement with the ESA team, which recommended focusing on such algorithms.

Destriping is the particular case of denoising in which N has a specific structure: it is directional and spatially coherent along one axis, appearing as a pattern of brighter and darker lines. The notion of a *stripe* is intrinsically two-dimensional and it does not have any meaning in a one-dimensional domain, which is why the destriping literature draws its methods predominantly from image processing. The MWI, however, does not acquire an image directly: it produces a continuous one-dimensional stream of measurements whose temporal correlation is the physical origin of the artefact, and a two-dimensional image is obtained only once consecutive acquisitions are stacked following the scanning geometry. The same phenomenon is therefore a correlated one-dimensional signal at the level of acquisition and a two-dimensional stripe pattern in the final product. Both readings apply for the MWI and algorithms in both forms will be tested.

The distinction between destriping and denoising also matters for evaluation: a method may lower the overall noise level without specifically addressing the striping pattern. While both are desirable, the focus should be placed on the latter, as residual uncorrelated noise is comparatively less harmful and can be reduced by simpler means. This duality is reflected in the metrics used to assess the selected methods in Chapter 6.

2.2 Striping in Satellite Instruments

Striping is a pervasive phenomenon in remote sensing imagery and its physical cause depends on the instrument architecture.

For multi-detector optical and infrared instruments the dominant cause is the difference in the response between detectors, which should instead respond consistently to the same received signal [12]. Bouali and Ladjal [13], analysing MODIS, distinguish three stripe types: detector-to-detector periodic stripes, mirror-side stripes, and random stripes confined to single thermal bands.

This leads to a distinction in the causes of striping:

- *Inter-detector* striping. Arises from systematic differences between the multiple detectors of an instrument and produces a fixed, typically periodic, spatial pattern tied to the detector layout.
- *Intra-detector* striping. Arises within a single detector from time-varying behaviour – nonlinear response drift, random fluctuations, low-frequency receiver noise – and produces a non-periodic, scene-independent pattern that no inter-detector calibration can remove.

The microwave instruments most relevant to the MWI fall in the second category. The total-power microwave radiometer used in such instruments uses a single receiver per channel, so there are no multiple detectors to compare against; striping is caused by low-frequency *flicker* noise originating in the receiver electronics. This mechanism has been investigated across the heritage humidity sounders: Qin et al. [14] trace the cross-track striping of the Advanced Technology Microwave Sounder (ATMS) Earth-scene observations to this low-frequency receiver noise, just as Dong and Zou do for the Tropical Rainfall Measuring Mission (TRMM) Microwave Imager (TMI) [15] and its successor Global Precipitation Measurement (GPM) Microwave Imager (GMI) [16].

In all of these conically or cross-track scanning radiometers the striping is non-periodic, scene-independent and rooted in the temporal correlation of a single receiver’s noise.

2.3 Destriping Methods

The literature of destriping methods can be grouped into four families [17]. Each is described below together with an assessment of its relevance to the MWI.

Statistical Methods

Statistical methods correct striping by aligning each detector’s statistical properties to a common reference. The two representatives are *histogram matching* and *moment matching*: the former relies on the assumption that the frequency histogram of pixel-values occurrences should show comparable profiles between detectors [12]; the latter assumes that detectors differ only in gain and offset and that they are related linearly [18].

These methods are designed for inter-detector striping and they are therefore inapplicable in the case of the MWI.

Filtering-based Methods

Filtering-based methods treat noise as energy concentrated in a characteristic region of a transform domain and suppress it there. A general limitation of the family is that signal structures sharing the same frequencies as the noise are attenuated along with it, leading to blurring and ringing artefacts [17].

The classical representative is the *Wiener Filter*, the linear minimum-mean-square-error filter when the signal and noise spectra are known. Operating directly on the measured PSD, it is the natural noise-aware, one-dimensional method and it is retained as such.

Extending the transform-domain approach to image denoising, one of the most

popular methods is *BM3D* [19], a two-stage non-local collaborative filtering: it groups mutually similar image patches into 3D groups and denoises them jointly through Wiener-type shrinkage in a sparsifying transform domain, exploiting non-local self-similarity. While the original formulation assumes white noise, its correlated-noise variant [20] accepts a known noise PSD, making it equally noise-aware.

As a state-of-the-art, two-dimensional denoiser, BM3D is retained as the second filtering method, complementing the one-dimensional Wiener filter.

Model-Optimization-based Methods

Model-Optimization-based methods solve the ill-posed problem of Equation 2.1 by minimising an energy functional combining a data-fidelity term with one or more regularisation terms that encode prior knowledge of the noise and the scene [13, 17]. The techniques of this family can be divided based on the prior they assume. *Variational* models constrain the recovered scene to be smooth or piecewise smooth; as such a prior can lead to blurring, *total-variation* methods instead impose directional smoothness, penalising the gradient across the stripe direction while preserving the one along it. The unidirectional variational model of Bouali and Ladjal [13] removes striping while preserving edges and scene structure. *Sparsity* and *low-rank* methods act on the stripe component itself, exploiting the low-rank attributes and the sparse properties of the gradient matrix in the stripe’s direction [21].

This last approach connects to the *decomposition* methods that constitute the operational benchmark for heritage microwave instruments. A swath dominated by striping is approximately low-rank, since the stripe offset is shared across a scanline and so contributes only to a few dominant modes.

Principal Component Analysis (PCA) exploits the same low-rankness directly: it concentrates the highest along-track variance, which corresponds to the striping pattern, into a few leading components. Combining PCA with *Ensemble Empirical Mode Decomposition* (EEMD), which adaptively separates a signal into intrinsic oscillatory modes, the high-frequency variations can be removed from the leading *Principal Component* (PC) coefficients while retaining the scene structure.

This *PCA/EEMD* strategy is the established destriping approach for instruments closely related to the MWI: it was applied to the ATMS by Qin et al. [14], treating only the first PC coefficient, and to the GMI by Dong and Zou [16], treating the first three coefficients but removing fewer modes from each.

It can be considered a 1D-in-2D hybrid – the decomposition is two-dimensional (PCA on the image) but the denoising acts on the one-dimensional along-track coefficient of each mode – and, as the operational benchmark for MWI-like instruments, it is retained.

Learning-based Methods

Learning-based methods use the same approach of Model-Optimization-based methods, but instead of relying on iterative algorithms they exploit deep-learning techniques, based mainly on Convolutional Neural Networks (CNNs). Although the literature around these methods has grown rapidly, achieving outstanding performance, their requirement is demanding in both the quantity and variety of data.

Given the very limited amount of MWI noise data, these methods are not considered. The effort of this work nonetheless points in their direction: the synthetic noise generators developed in Chapter 4 are used to produce exactly such a dataset, by injecting validated synthetic noise into clean Earth scenes, and the purpose of the scientific challenge is to allow the research community to explore learning-based destriping.

In summary, the three methods selected to span the destriping domain are:

- *Wiener Filter* (Section 5.1): a one-dimensional, frequency-domain filter, exploiting the noise PSD directly.
- *PCA/EEMD* (Section 5.2): the operational benchmark for heritage microwave imagers, exploiting the low-rank mode structure of the striped swath.
- *BM3D* (Section 5.3): a state-of-the-art image processing denoiser, whose correlated-noise variant accepts a known noise PSD.

To these is added a fourth, hybrid method, combining PCA/EEMD and BM3D, whose implementation is motivated in Section 5.4.

Chapter 3

Instrument Noise Modelling

This chapter conducts a statistical analysis of the MWI instrument noise from ground-test data. The analysis proceeds in two stages: first, the noise is studied and characterised as a one-dimensional *time series*; then, it is examined in its two-dimensional *spatial* form, formally introducing the *striping* feature.

3.1 Time Series Analysis of Instrument Noise

The instrument noise is first analysed as a temporal signal, which is how it is acquired, before considering any spatial structure. The objective is to define the characterising properties of the noise, later used to develop and validate the synthetic generators.

3.1.1 Ground-Testing Setup and Data

Several pre-flight test campaigns were carried out on the MWI instrument, producing datasets that differ in target type, duration and scanning configuration. The dataset made available for this study is the MWI-13 V receiver raw dBm data: continuous measurements acquired without the scan rotation by simply pointing at a target at fixed temperature; this test allows to capture the intrinsic instrument noise, isolating it from the scanning geometry.

Each acquisition is labelled by two settings: the receiver temperature $T_{rec} \in \{-6, 22, 36\}^\circ\text{C}$ and the target type $T_{tar} \in \{Meas, Hot, Cold\}$. All three receiver temperatures are analysed, although the noise model is ultimately based on the operational setting $T_{rec} = 22^\circ\text{C}$. The three target types correspond to three distinct constant target temperatures: *Cold* and *Hot* are the cold and hot references, while

Table 3.1: Summary of the available ground-test acquisitions.

T_{rec} [°C]	Date	<i>Meas</i>	<i>Hot</i>	<i>Cold</i>
−6	2021-05-12	5	1	1
22	2021-05-07	5	1	1
	2021-05-11	5	1	1
	2021-09-06	1	0	0
36	2021-05-11	5	1	1
	2021-09-06	5	1	1
Total		26	5	5

Meas is an intermediate Earth-like target.

The measurements were collected over four days between May and September 2021. Within a session, the targets are recorded at the sampling time of 1 ms and in a fixed order: five *Meas* acquisitions of approximately 200 s ($\sim 200\,000$ samples), followed by one *Hot* and one *Cold* reference of about 10 s each ($\sim 10\,000$ samples). The five *Meas* acquisitions are recorded essentially contiguously, about 30 ms apart, while the *Hot* and *Cold* references follow after a short gap of roughly 20 s.

The full dataset is summarized in Table 3.1; note that not every setting combination (T_{rec} , T_{tar}) is available for every acquisition date and the 2021-09-06 session contains a single (22 °C, *Meas*) measurement with no *Hot* and *Cold* references.

In total, the dataset comprises 26 *Meas*, 5 *Hot* and 5 *Cold* acquisitions.

3.1.2 First and Second Order Statistics

The mean and standard deviation of each acquisition are examined to characterise the measured signal and how the noise depends on the operating conditions. Each acquisition is first preprocessed by removing samples beyond $\pm 4\sigma$ of the mean as outliers and converting the raw values from dBm to power in Watts.

Mean and Standard Deviation

The dataset consists of acquisitions taken in ambient rather than TVAC testing, so it is first verified that the room temperature variations are negligible, so that the target contribution can be treated as a constant offset; indeed, across all settings with repeated acquisitions, the coefficient of variation of the mean stays within the range 0.03%–0.12%.

The standard deviation, unlike the mean, is not stable. Two distinct sources of variability are observed:

Table 3.2: Mean and coefficient of variation (CV) of the per-acquisition noise standard deviation [W], aggregated by target type and receiver temperature.

T_{tar}	T_{rec} [°C]	Mean std [W]	CV[%]
<i>Cold</i>	−6	1.458×10^{-8}	-
	22	1.938×10^{-8}	5.7
	36	2.027×10^{-8}	0.5
<i>Hot</i>	−6	2.051×10^{-8}	-
	22	2.778×10^{-8}	9.3
	36	3.184×10^{-8}	13.3
<i>Meas</i>	−6	2.624×10^{-8}	12.9
	22	3.472×10^{-8}	15.9
	36	3.847×10^{-8}	9.7

- *Across conditions.* The mean noise standard deviation increases systematically with both target and receiver temperature (Table 3.2), which is physically expected as the noise amplitude scales with the total system temperature.
- *Within a fixed condition.* Even under identical experimental conditions (i.e. same receiver temperature and target type) the standard deviation is not stable across acquisitions. Considering the 11 measurements in the setting (22 °C, *Meas*), the standard deviation ranges from 2.88×10^{-8} to 4.35×10^{-8} , with no monotonic trend within a day.

Because the noise amplitude varies with the operating conditions, the characterisation is carried out at a single fixed operating point, while scaling to arbitrary scene temperatures is addressed separately in Section 3.2.2. The dataset used for the remainder of the analysis is therefore restricted to the operational receiver setting $T_{rec} = 22$ °C and target type $T_{tar} = Meas$, as the *Hot* and *Cold* references are both too few (a single acquisition per session) and too short (10 s each) to support a reliable characterisation. This leaves the 11 (22 °C, *Meas*) acquisitions shown as a concatenated time series in Figure 3.1.

The within-setting variability, on the other hand, is far larger than finite-sample estimation theory would predict; the observed max-to-min variance ratio reaches 2.28, whereas for an independent and stationary process the maximum expected ratio is 1.02 (at 95% confidence), motivating the investigation of the stationarity and correlation structure of the noise that follows.

The marginal distribution is approximately Gaussian: skewness and excess kurtosis remain close to zero, with 88.89% of the acquisitions having both below 0.2, as confirmed by the quantile-quantile plots.

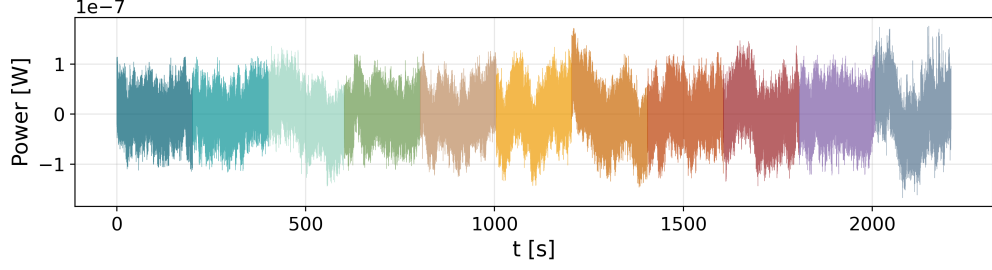


Figure 3.1: Acquisitions in the setting (22 °C, *Meas*) concatenated into a single zero-mean noise time series, with each colour denoting one acquisition.

Stationarity: ADF and KPSS tests

Before formally describing the tests, the two standard notions of stationarity for a stochastic process $\{X_t\}$ are recalled.

Definition 3.1 (Strict-sense stationarity). A discrete-time stochastic process $\{X_t\}$ is said to be *strict-sense stationary* (SSS) if for all $n \in \mathbb{N}^+$, $(t_1, t_2, \dots, t_n) \in T^n$ and for every choice of h such that $t_i + h \in T$, for all $i = 1, \dots, n$ it holds

$$P(X_{t_1} = x_1, \dots, X_{t_n} = x_n) = P(X_{t_1+h} = x_1, \dots, X_{t_n+h} = x_n).$$

Definition 3.2 (Wide-sense stationarity). A discrete-time stochastic process $\{X_t\}$ is *wide-sense stationary* (WSS) of order 2 if for all $t_i, t_i + l, t_i + l + h \in T$

$$\begin{aligned} E(X_{t_i}) &= E(X_{t_i+h}), \\ E(X_{t_i}X_{t_i+l}) &= E(X_{t_i+h}X_{t_i+l+h}). \end{aligned}$$

As stationarity is a fundamental property for synthetic noise generation, both the Augmented Dickey–Fuller (ADF) and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) tests are applied [22]:

- *ADF test*. Tests the null hypothesis of a unit root in the series against the alternative of stationarity:

$$\begin{aligned} H_0 &: \text{The series has a unit root (non-stationary),} \\ H_1 &: \text{The series has no unit root.} \end{aligned}$$

- *KPSS test*. Tests the null hypothesis of stationarity against the alternative of the presence of a unit root:

$$\begin{aligned} H_0 &: \text{The process is stationary,} \\ H_1 &: \text{The series has a unit root (non-stationary).} \end{aligned}$$

Table 3.3: Subsample stationarity analysis for two representative acquisitions: median p-values and rejection rates of the ADF and KPSS tests.

Duration [s]	Low-variance acquisition				High-variance acquisition			
	ADF		KPSS		ADF		KPSS	
	p	rej.rate	p	rej.rate	p	rej.rate	p	rej.rate
5	0.168	0.30	0.050	0.50	0.088	0.30	0.056	0.50
10	0.063	0.50	0.018	0.60	0.031	0.60	0.032	0.60
15	0.005	0.80	0.049	0.50	0.018	0.75	0.010	0.95
20	0.006	0.75	0.010	0.75	0.006	0.80	0.010	0.85
25	0.002	1.0	0.010	0.90	0.003	1.0	0.010	0.90
30	0.003	1.0	0.010	0.90	0.005	0.95	0.010	1.0
35	0.001	1.0	0.010	1.0	0.002	0.95	0.010	1.0
40	0.000	1.0	0.010	0.95	0.001	1.0	0.010	1.0

Having opposite null hypotheses, the joint application of the two tests can lead to two cases: either the tests agree (both supporting or rejecting stationarity), or one supports stationarity while the other rejects it. For all acquisitions the tests fall in the latter: ADF rejects the unit-root null while KPSS rejects stationarity, both with p-values well below $p = 0.05$.

Given the acquisition length ($N \sim 200\,000$), the tests are repeated on shorter subsamples to verify that the result is not a large-sample artefact. For each target length, 20 subsamples are drawn from random start positions and the rejection rate is taken as the fraction rejecting at $p = 0.05$ ¹(Table 3.3). The result is confirmed: whether the rejection rates are low (shorter duration) or high (longer duration), the two tests keep giving opposite verdicts on the series.

This simultaneous rejection rules out both a purely stationary process (ADF rejects, KPSS does not) and a unit-root process (ADF does not reject, KPSS does), a behaviour that can arise under long-range dependence. Two observations point in this direction: the order in which the two tests reach high rejection depends on the acquisition – in the low-variance one the ADF rejects the unit root earlier, in the high-variance one the KPSS rejects stationarity earlier; furthermore, the outcomes proved sensitive to the lag truncation, leading to the choice of a high fixed lag value, without affecting the conclusion.

The nature of this behaviour is clarified in the following section, where the correlation and time-frequency structure of the noise are analysed directly.

¹The KPSS implementation returns p-values from a lookup table; when the test statistic exceeds the tabulated range, the reported value is capped at 0.01 and the true p-value is smaller.

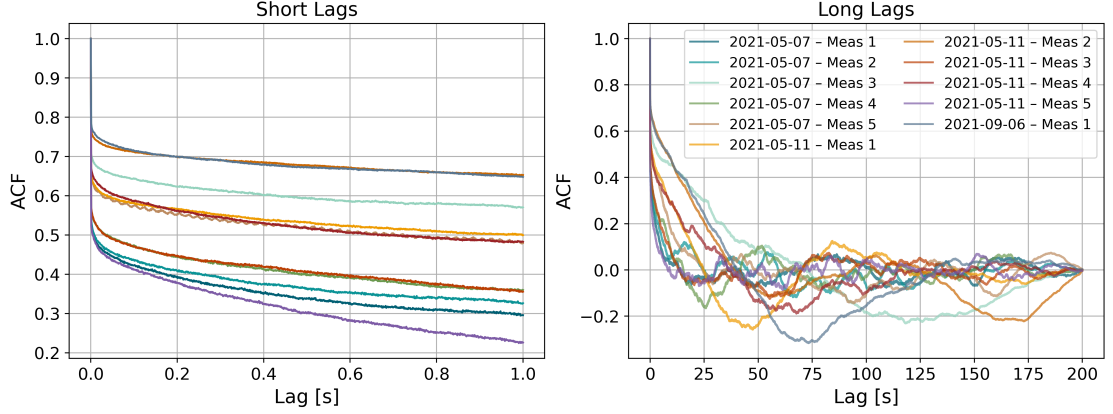


Figure 3.2: Autocorrelation function of the (22 °C, *Meas*) acquisitions. *Left:* short lags (up to 1 s), showing the initial decay. *Right:* long lags (up to 200 s), revealing the zero-crossing and the negative lobes.

3.1.3 ACF, Allan Deviation and PSD

In this section the noise is analysed with three complementary tools: the autocorrelation function (ACF), the Allan deviation and the power spectral density (PSD).

Already in the time series of Figure 3.1, two coexisting behaviours can be recognized: a fast fluctuation from sample to sample and a slower wandering of the local level that drifts over several seconds. The following analysis formalizes this qualitative picture, showing that the noise is the sum of a fast, high-frequency component and a strongly correlated component dominating at lower frequencies.

Correlation Analysis: Autocorrelation Function

As shown in the plots of Figure 3.2, the correlation structure of the noise is consistent across acquisitions, while individual realizations show variability. Three features characterise the behaviour: a strong lag-1 correlation ($[0.53, 0.78]$); a slow, power-law decay over several seconds, taking from 1.3 to 32.3 s to fall below 0.2; a finite correlation horizon, with the zero-crossing of tens of seconds ($[8, 72]$ s), followed by negative lobes. The *Hot* and *Cold* references, although too short (10 s) for a reliable estimate, exhibit the same qualitative behaviour.

The power-law decay, confirmed by a slowly decreasing partial autocorrelation function (PACF), rules out any finite-order $AR(p)$ model, and the finite correlation horizon and negative lobes rule out a random walk, consistent with the rejection of the unit-root null by the ADF test. The process is indeed subject to long-range correlations, which produce sign-changing excursions that span a wide range of

durations, with a characteristic timescale of 40 s on average. The long memory is quantified by fitting a fractionally-integrated ARFIMA(1, d , 0) model, which yields stable estimates across all acquisitions with $d \approx 0.55 \pm 0.017$. The order of differencing d both confirms strong long-range dependence ($d > 0$) and places the process at the edge of stationarity ($d = 0.5$), consistent with the ADF and KPSS results. The ARFIMA model is used only to characterise this behaviour: being a single-component process, it cannot reproduce both the white fast fluctuations and the long-memory slower drifts, which is the role of the two-component model developed in Chapter 4.

Finally, the variability across the ACF curves is shown to be intrinsic to the process rather than experimental: two same-day acquisitions can differ markedly, while acquisitions from different days can be nearly identical. The curves with higher short-lag correlation and longer correlation horizon are precisely those with higher total variance, all driven by the specific realization of the long memory. When it is realized more strongly, it causes larger and longer excursions, simultaneously raising the lag-1 correlation, the correlation horizon and the total variance. This coupling is confirmed by strong rank correlations between the standard deviation and the lag-1 ACF ($\rho = 0.91$), the standard deviation and the zero-crossing time ($\rho = 0.76$), and the lag-1 ACF and the zero-crossing time ($\rho = 0.88$).

Component Analysis: Allan Deviation

The Allan variance is defined as follows:

Definition 3.3 (Allan variance). Given a series sampled at interval τ_0 , let $\bar{y}_i(\tau)$ denote the average of the signal over the i -th window of length $\tau = m\tau_0$. The Allan variance is half the mean square of the differences between successive window averages

$$\sigma_y^2(\tau) = \frac{1}{2(M-1)} \sum_{i=1}^{M-1} \left(\bar{y}_{i+1}(\tau) - \bar{y}_i(\tau) \right)^2.$$

The Allan deviation (hereafter also ADEV) is the square root of the Allan variance and is used to identify the components of the noise, as different noise types produce characteristic slopes in its curve.

As shown in Figures 3.3 and 3.4, all acquisitions share the same structure, with a local minimum at $\tau \sim 0.07$ s separating two regimes. At short averaging times the fitted slope is -0.45 , close to the -0.5 of pure *white noise*: averaging uncorrelated samples over a longer window reduces their fluctuation, so the deviation between successive windows decreases. Past the minimum the curve rises, as the second component takes over: the *flicker* (or $1/f$) noise, a process whose energy dominates

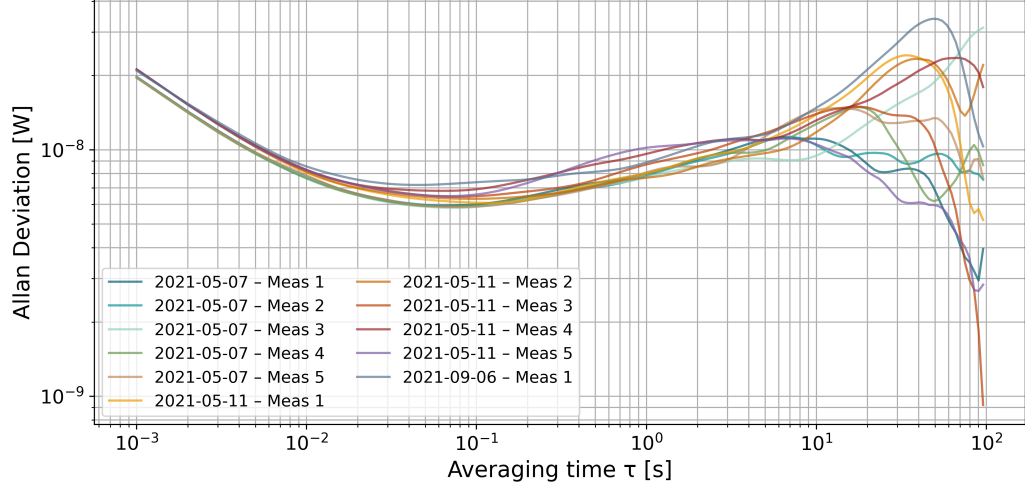


Figure 3.3: Allan deviation of the (22 °C, *Meas*) acquisitions. The curves share a consistent shape with a minimum at $\tau \sim 0.07$ s and a peak near $\tau \sim 40$ s.

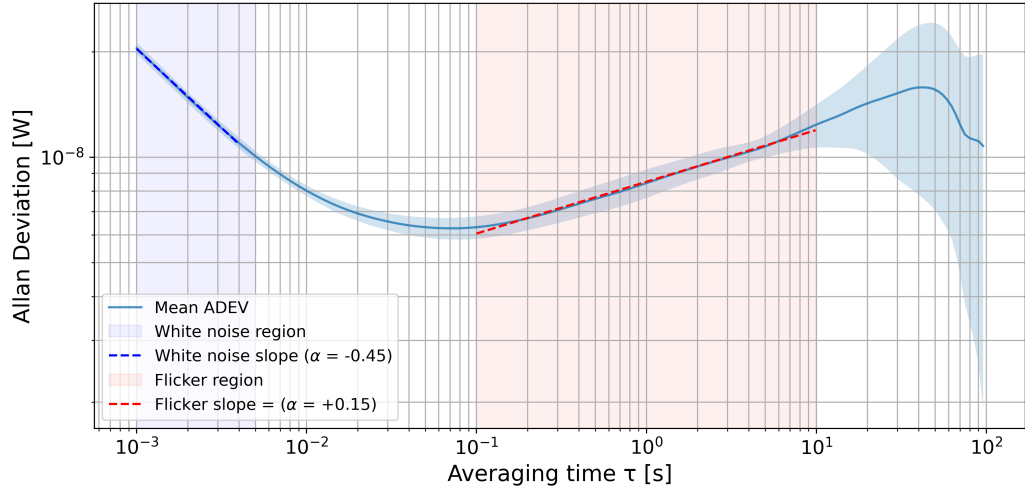


Figure 3.4: Mean Allan deviation with variability band and fitted slopes in the white noise and the flicker-dominated regions.

at lower frequencies, causing slower drifts of the local level that do not average out, so successive window means differ more as the window grows. The fitted slope in this region is $+0.15$, close to the 0 of ideal flicker noise and further from the $+0.5$ of a random walk, consistent with $1/f$ behaviour rather than an unbounded drift. The mean ADEV peaks at $\tau \sim 40$ s, in agreement with the mean ACF zero-crossing, and then turns down. This peak marks the characteristic timescale of the flicker

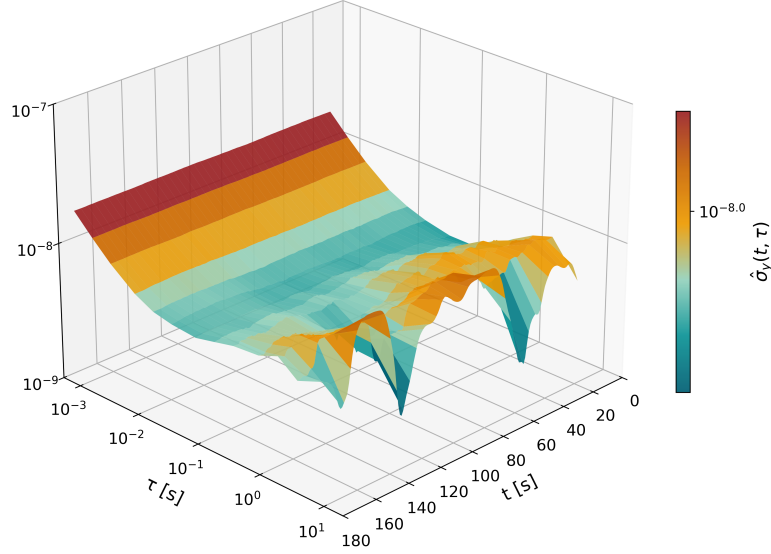


Figure 3.5: Dynamic Allan deviation for a representative (22 °C, *Meas*) acquisition. Computed using a sliding window of length 40 s advanced in steps of 1 s. Within each window only the reliable estimates are retained, set to $\tau_{\max} = 40/3 \text{ s} \approx 13.3 \text{ s}$.

excursion: it is the averaging time at which successive window means are most dissimilar, i.e. one window tends to sit above the local level and the next below it. For longer windows each average spans several excursions, so successive means converge again and the deviation falls. The larger variability band around the peak (Figure 3.4) mirrors the wide spread of the ACF zero-crossings.

Beyond $\sim 50 \text{ s}$ the estimator becomes unreliable due to the limited number of independent windows: the reliability limit of the ADEV is empirically set to $N/4$, where N is the series length. This is a limitation of the test setup – longer continuous acquisitions would allow to characterise the flicker component and its peak with confidence. Since the behaviour is consistent across acquisitions, the concatenated series is used to extend the reliable range, confirming the result.

The variability caused by the flicker component can be visualized directly through the dynamic Allan deviation [23], which evaluates the ADEV over a sliding window, producing a surface as a function of both averaging time τ and position t along the series (Figure 3.5). The surface is smooth at short τ , where the white component dominates, but becomes increasingly irregular at long τ : there each finite window captures only part of a single flicker excursion, so the estimate reflects the particular low-frequency content of that window rather than a stable ensemble value.

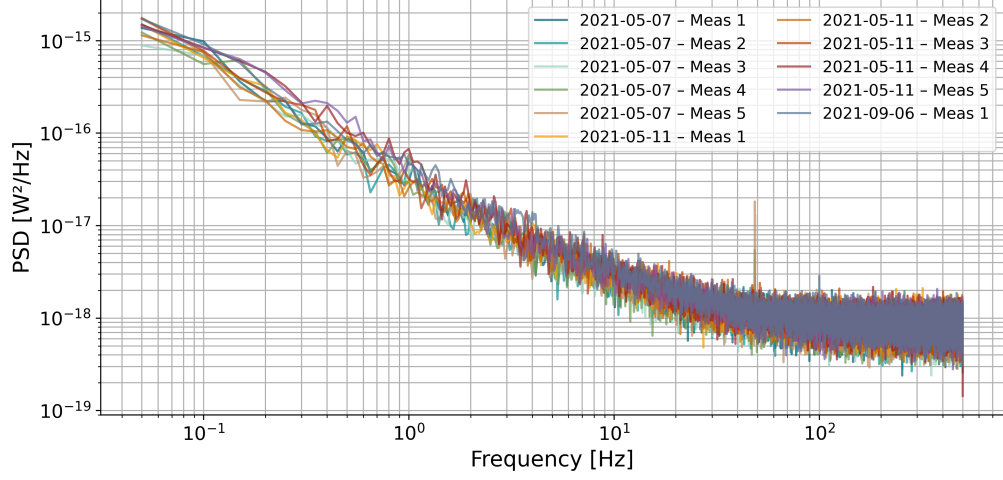


Figure 3.6: Power spectral density of all the (22 °C, *Meas*) acquisitions. The individual Welch estimates, each obtained by averaging the periodograms of overlapping 20 s segments, show a highly reproducible spectral shape.

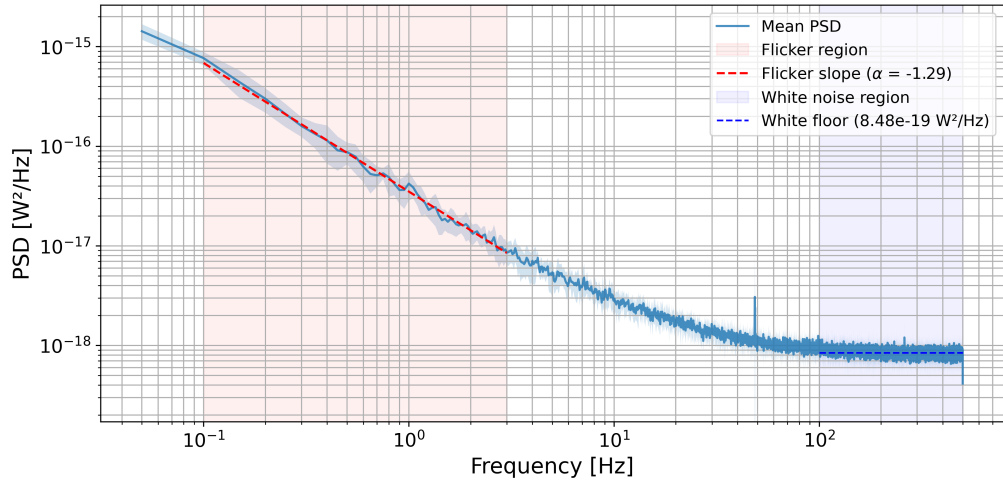


Figure 3.7: Mean power spectral density with variability band and fitted slopes in the white noise and the flicker-dominated regions.

Frequency Analysis: Power Spectral Density

The two-component structure is confirmed directly in the frequency domain by the power spectral density (PSD), estimated using Welch's method, which averages periodograms over overlapping segments to reduce variance [24]. The spectrum has a flat thermal floor at high frequencies, fitted at $8.48 \times 10^{-19} \text{ W}^2/\text{Hz}$ above 100 Hz,

and a power-law rise towards low frequencies with slope $\alpha = -1.29$ (Figure 3.7). The slope is slightly steeper than the -1 of ideal flicker noise and further from the -2 of a random walk, reflecting the behaviour observed in the Allan deviation.

The lowest accessible frequency is set by the acquisition length: for a single 200 s measurement it is $1/200 \text{ s} = 0.005 \text{ Hz}$, but resolving it in a single segment comes at the cost of a high-variance estimate, which is why the Welch average trades resolution for reliability (in Figure 3.6 the lowest frequency is indeed $1/20 \text{ s} = 0.05 \text{ Hz}$). As noted for the Allan deviation, longer measurements would both improve the estimate and resolve the spectrum at lower frequencies. In this case, this would allow the low-frequency cut-off to be determined: below it the $1/f$ rise flattens, bounding the total power contribution of the flicker component.

Figures 3.6 and 3.7 provide crucial information for the noise model: the spectral shape is highly reproducible across acquisitions as the curves overlie one another closely, confirmed by the tight variability band around the mean PSD. While the finite-window time-domain statistics fluctuate from one acquisition to the next, the underlying ensemble spectral structure is stable.

3.1.4 Noise Model

The information extracted from the analysis can now be unified into a model of the measured signal. Each acquisition is described as

$$x(t) = \mu + n(t), \quad (3.1)$$

where μ is a constant offset and $n(t)$ is a zero-mean stochastic noise process. The offset μ collects both the constant target contribution and the day-to-day experimental setup conditions; both are removed in preprocessing, leaving $n(t)$ as the object of the model. The characterisation established three properties of $n(t)$.

Two-component structure. The noise is the sum of a white and a flicker-like component, dominant respectively below and above $\tau \sim 0.07 \text{ s}$ in the time domain, and corresponding to the plateau and the $1/f$ -like rise at higher and lower frequencies in the spectral domain.

Long memory. The flicker component causes long-range temporal correlations, confirmed by a fractional integration order $d > 0$ and characterised by a finite correlation horizon followed by negative lobes. This produces excursions of the local level whose characteristic timescale – set by the ACF zero-crossing and the ADEV peak – is approximately 40 s on average, though broadly distributed across acquisitions (median $\approx 25 \text{ s}$).

Wide-sense stationarity. A pure $1/f$ process is, strictly speaking, non-stationary, as its variance diverges with the unbounded low-frequency content. In real physical systems, however, the variance is bounded either by an intrinsic low-frequency cut-off or by the fact that observing the process over a finite window resolves no frequency below $\sim 1/N$; both mechanisms justify treating the process as approximately stationary.

Despite the conflicting result of the ADF and KPSS tests, several elements support the stationarity assumption: a finite correlation horizon, the reversion of the flicker excursion to the mean and a reproducible spectral structure; in a pure non-stationary process these properties would change substantially between measurements.

For modelling purposes, the noise is therefore treated as wide-sense stationary, with the understanding that this holds as an approximation for a process that, with $d \approx 0.55$, lies just past the boundary between stationarity and non-stationarity. For a Gaussian process, wide-sense stationarity would imply strict-sense stationarity; however, since neither the Gaussianity nor the wide-sense stationarity holds exactly, this stronger conclusion is not claimed.

This model – a Gaussian, two-component, long-memory noise with a reproducible power spectral density, treated as wide-sense stationary – is what the synthetic noise generators of Chapter 4 must reproduce.

3.2 Spatial Manifestation of Noise: Striping

The acquired one-dimensional signal is mapped onto a two-dimensional swath image by the scanning geometry introduced in Section 1.2, producing the striping pattern. This section first quantifies the relative weight of the two noise components, then describes the operational calibration through which striping is partly mitigated, and finally analyses statistically how often striping occurs and whether it can be detected operationally.

3.2.1 Striping Metrics and Quantification

Three metrics are used throughout this section to quantify striping.

Striping Index. The first metric operates directly in the image domain, as defined by Qin et al. [14]. Based on the directional decomposition of the noise (Section 1.2), striping is a coherent offset shared along the scan direction, so it inflates the along-track variance relative to the cross-track one.

The *Striping Index* (SI) is defined as the ratio of the mean along-track variance V_{along} to the mean cross-track variance V_{cross}

$$\text{SI} = \frac{V_{\text{along}}}{V_{\text{cross}}} \quad (3.2)$$

where, denoting by $T_b(i, j)$ the brightness temperature at the i -th scanline ($i = 1, \dots, L$) and j -th cross-track sample ($j = 1, \dots, M$), the two variances are

$$V_{\text{along}} = \frac{1}{M} \sum_{j=1}^M \left(\frac{1}{L} \sum_{i=1}^L \left(T_b(i, j) - \frac{1}{L} \sum_{i=1}^L T_b(i, j) \right)^2 \right), \quad (3.3)$$

$$V_{\text{cross}} = \frac{1}{L} \sum_{i=1}^L \left(\frac{1}{M} \sum_{j=1}^M \left(T_b(i, j) - \frac{1}{M} \sum_{j=1}^M T_b(i, j) \right)^2 \right). \quad (3.4)$$

A value of $\text{SI} \approx 1$ indicates the absence of striping, being the along-track and cross-track variances comparable; increasingly larger values correspond to a progressively stronger coherent along-track structure.

Striping Percentage and Ratio. The remaining two metrics quantify the relative contribution of the thermal and flicker components, following the procedure of Yang et al. [10]. The magnitude of the noise is measured using the noise equivalent delta temperature (NEDT), defined as the square root of the integral of the power spectral density

$$\text{NEDT}_{\text{total}} = \sqrt{\int_{f_{dn}}^{f_{up}} S(f) df},$$

or, equivalently, as the unbiased second central moment of noise in the time domain

$$\text{NEDT}_{\text{total}} = \sqrt{\frac{1}{LM-1} \sum_{i=1}^L \sum_{j=1}^M (n_{i,j} - \bar{n})^2};$$

it is decomposed into a thermal and a $1/f$ component as

$$\text{NEDT}_{\text{total}}^2 = \text{NEDT}_{\text{thermal}}^2 + \text{NEDT}_{1/f}^2.$$

The thermal contribution can be isolated in either domain: in the frequency domain, it is obtained by integrating $S(f)$ above the corner frequency that separates the low- and high-frequency regimes; in the time domain, it is estimated from the variance of the differences between adjacent samples, which naturally captures the high-frequency noise. In both cases the $1/f$ component follows as the residual. From these quantities, the *Striping Percentage* ($P_{1/f}$) and *Striping Ratio* ($R_{1/f}$) metrics are defined as

$$P_{1/f} = \frac{\text{NEDT}_{1/f}^2}{\text{NEDT}_{\text{total}}^2} \times 100\%, \quad R_{1/f} = \frac{\text{NEDT}_{1/f}}{\text{NEDT}_{\text{thermal}}}. \quad (3.5)$$

Using the time-domain implementation on the concatenated time series (Figure 3.1), MWI-13 V noise results in a total NEDT of 0.964 K ($\sigma_{\text{thermal}} = 0.564$ K and $\sigma_{1/f} = 0.782$ K), a Striping Percentage of 65% and a Striping Ratio of 1.39; with the former above 50% and the latter above 1, the 21st channel of the MWI instrument firmly sits in the severe-striping regime.

Computing the metrics on each acquisition separately confirms the result: the relative flicker contribution ($P_{1/f}$) varies with the realization, as expected, but its concatenated value is well-centred within the per-acquisition distribution.

3.2.2 Operational Calibration Process

The available data are one-dimensional, fixed-target recordings, with no scanning and no calibration applied. To study how the on-orbit two-point calibration interacts with striping and, later, to inject the synthetic noise into simulated Earth scenes, both the scan rotation and the calibration must be reproduced. This in turn requires accounting for the signal dependence of the noise established in Section 3.1.2, since the *Cold*, *Hot* and *Earth* segments are observed at different temperatures.

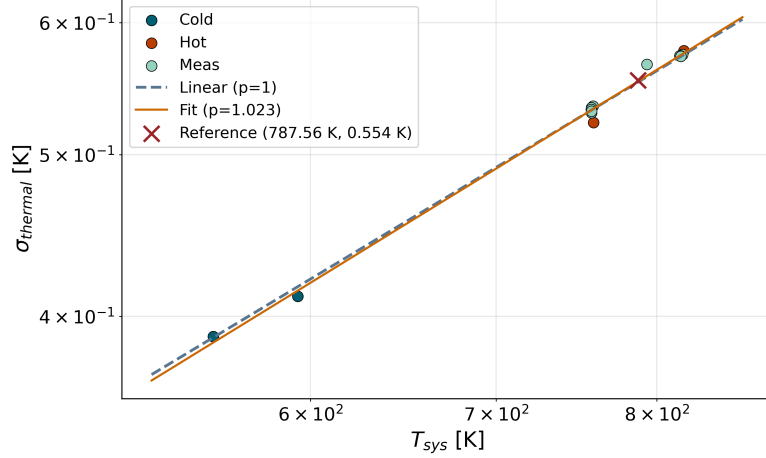


Figure 3.8: Validation of the linear noise scaling law. The red marker indicates the laboratory reference operating point (787.56 K, 0.554 K).

Noise Scaling

The sensitivity of a total-power radiometer [25] is

$$\Delta T = T_{sys} \cdot \sqrt{\frac{1}{B\delta} + \left(\frac{\Delta G}{G}\right)^2}, \quad T_{sys} = T_{n,rec} + T_{scene}$$

where B is the bandwidth, δ the integration time, $\Delta G/G$ the gain fluctuation and T_{sys} the total system temperature, defined as the sum of the receiver noise temperature $T_{n,rec}$ and the recorded scene temperature T_{scene} .

The thermal contribution gives a standard deviation of $\sigma_{thermal} = T_{sys}/\sqrt{B\delta}$, while the second term captures the $1/f$ contribution, which is also proportional to T_{sys} to first order. Both components therefore scale linearly with the total system temperature, so the noise measured at the laboratory reference point can be rescaled to any scene temperature through the thermal amplitude, which sets the reference:

$$\sigma_{thermal}(T_{scene}) = \sigma_{thermal,lab} \cdot \frac{T_{n,rec} + T_{scene}}{T_{sys,lab}}. \quad (3.6)$$

Only the amplitude changes, while the noise shape is preserved: this separability between instrument-dependent shape and scene-dependent amplitude is what allows the laboratory characterisation to generalize to orbit conditions.

The law is validated in Figure 3.8: fitting $\sigma_{thermal}(T_{tar}) \propto T_{sys}^p$ across the acquisitions of all three target types at $T_{rec} = 22$ °C yields $p = 1.02 \pm 0.02$, consistent with the linear scaling.

The laboratory reference values are obtained by averaging the $T_{tar} = Meas$ acquisitions, giving $\sigma_{thermal,lab} = 0.554$ K, $T_{sys,lab} = 787.56$ K.

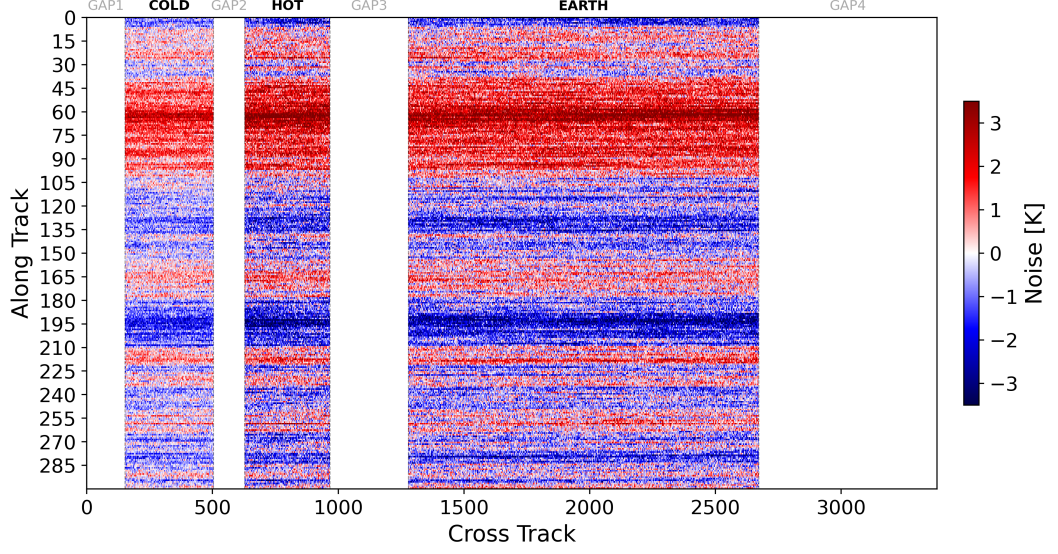


Figure 3.9: Reconstructed scan sequence: the measured noise folded into *Cold*, *Hot* and *Earth* segments with non-observation gaps discarded and each segment scaled to its operating temperature.

Noise Processing for Operational Use

The selected laboratory measurements in the setting (22 °C, *Meas*) are preprocessed into a single noise sequence through four steps:

1. *Demeaning.* The signal mean is removed from each acquisition so that only the zero-mean noise $n(t)$ enters the model (Section 3.1.2).
2. *Amplitude normalisation.* The white noise amplitude varies slightly across acquisitions – a standard deviation of 0.021 K around $\sigma_{thermal,lab} = 0.554$ K – so each series is normalised to $\sigma_{thermal,lab}$.
3. *Power-to-temperature conversion.* Each acquisition is converted from power in Watts P_W to temperature in Kelvin T through the inverse radiometer relation:

$$T = \frac{P_W - P_{off}}{G_{rx}} - T_{n,rec},$$

where $P_{off} = 1.09 \times 10^{-9}$ W, $G_{rx} = 3.68 \times 10^{-8}$ W/K and $T_{n,rec} = 500.254$ K are the given laboratory reference values. The processed acquisitions are then concatenated into a single continuous series; this concatenated sequence is also the dataset used to define the synthetic noise generators of Chapter 4, which produce noise already in Kelvin at the laboratory sampling rate and therefore enter the pipeline at step 4.

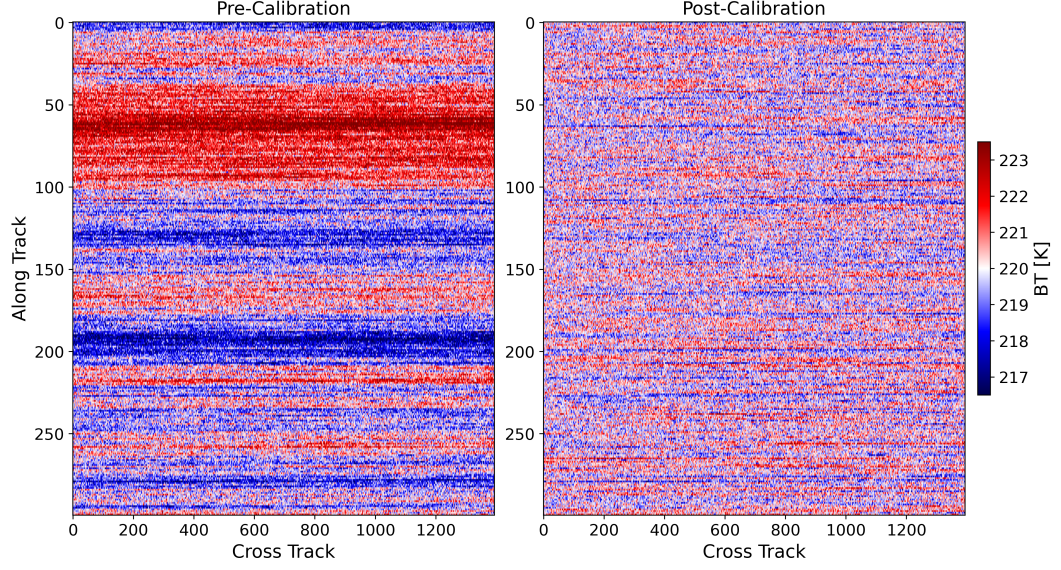


Figure 3.10: Spatial manifestation of striping, before (left) and after (right) two-point calibration, for a constant Earth scene at 220 K.

4. *Resampling.* The concatenated series is resampled from $T_{s,\text{lab}} = 1$ ms to the operational rate $T_{s,\text{op}} = 0.394$ ms. Because a shorter integration time increases the white noise amplitude, the series is rescaled by $\sqrt{T_{s,\text{lab}}/T_{s,\text{op}}}$, preserving the relative flicker contribution.

The resulting series is a zero-mean, temperature-domain noise sequence at the operational sampling rate, ready to be folded into the scan geometry and rescaled to the operating temperature of each segment, as described in the next section.

Calibration Simulation

The processed noise sequence is folded into the scan geometry (Figure 3.9): each antenna rotation is partitioned into *Cold*, *Hot* and *Earth* segments separated by non-observation gaps, whose samples are discarded, and each active segment is rescaled to its physical temperature through Equation 3.6.

Calibration is the linear interpolation that maps Earth counts C_e to temperature T_e using the cold and warm target values (C_c, T_c) , (C_w, T_w) :

$$T_e = \frac{T_w - T_c}{C_w - C_c} \cdot (C_e - C_w) + T_w,$$

with $C = (T_{scene} + T_{n,rec}) \cdot G$. Although the data are not available in counts, the gain G cancels in the ratio, so the procedure can be reproduced entirely in the temperature domain.

Figure 3.10 shows the spatial effect of striping: the simulated *Earth* scene is held at a constant arbitrary brightness temperature of $T_{scene} = 220$ K, while the warm and cold references at $T_w = 282.05$ K (see Section 6.1.1) and $T_c = 2.73$ K (deep space). Pre-calibration, the flicker drifts appear as coherent horizontal bands along the scan direction, confirming that the long-range temporal correlation characterised in Section 3.1.3 manifests two-dimensionally as striping. After calibration the broad bands (i.e. flicker excursions spanning several scan rotations) are largely removed, but finer residual stripes (i.e. flicker excursions shorter than a scan rotation) persist, illustrating that calibration attenuates but doesn't resolve the striping problem.

3.2.3 Statistical Analysis of Stripe Occurrence

The following analysis examines how often striping actually occurs, how effectively calibration suppresses it and whether it can be detected from operationally available quantities. Since this requires control over the noise composition – isolating the flicker contribution in some stages and using the full noise in others – the study is carried out on synthetic Thermal+1/ f noise, generated as described in Chapter 4. The scan and calibration are reproduced as in the previous section, and the Earth scene is held at a constant reference temperature of 220 K, so that any spatial structure is attributable to noise alone.

The analysis is run over 4545 scanlines, which is approximately one full orbit.

Performing the study first requires an operational definition of a *stripe*: a coherent offset shared by the Earth samples of a scanline, as opposed to the sample-to-sample scatter of thermal noise. The detection and classification proceeds per scanline, in three steps:

1. Each Earth scanline is divided into $n_{seg} = 8$ sub-windows of ~ 175 samples each; averaging within a sub-window cancels the incoherent thermal scatter, leaving only a coherent offset if one is present.
2. Each sub-window is labelled according to its mean error relative to the reference scene temperature: +1 if the error exceeds $+\tau$, -1 if below $-\tau$ and 0 otherwise, where $\tau \in \{0.3, 0.7, 1.0\}$ K is the error threshold.

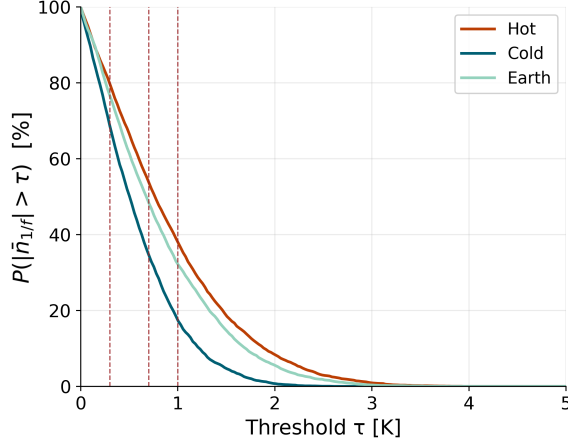


Figure 3.11: Probability that the per-scanline flicker mean exceeds a threshold τ , for the *Cold*, *Hot* and *Earth* windows. Dashed lines mark the thresholds $\tau = 0.3, 0.7, 1.0$ K.

3. Each scanline is then classified as:

- *Full*. At least 6 of the 8 sub-windows share the same nonzero sign, indicating a stripe spanning the whole scanline.
- *Partial*. If not full but at least 2 adjacent sub-windows share the same sign label, indicating a localized stripe artefact.
- *Clean*. Not full nor partial.

Flicker study

This first stage examines the flicker component, the direct cause of striping.

The probability that the per-scanline flicker mean exceeds a threshold τ in magnitude, $P(|\bar{n}_{1/f}| > \tau)$ – the drift being equally likely above or below the true signal – is computed for the *Cold*, *Hot* and *Earth* windows (Figure 3.11), and their inter-window correlation is examined.

Since the noise amplitude scales with the target temperature, the *Cold* and *Hot* windows bracket the *Earth* view occurrence from below and above: at $\tau = 1$ K the flicker noise exceeds the threshold in 17.5% of scanlines for the *Cold* window and 38% for the *Hot* window, with the *Earth* view at 33% in between. The occurrence drops steeply with the threshold, falling from 69–80% at 0.3 K to below 10% for all windows beyond 2 K.

The flicker component is strongly correlated across the three windows at the scanline level, with Pearson coefficients of $r = 0.87$ (*Hot–Earth*), $r = 0.84$ (*Cold–Earth*)

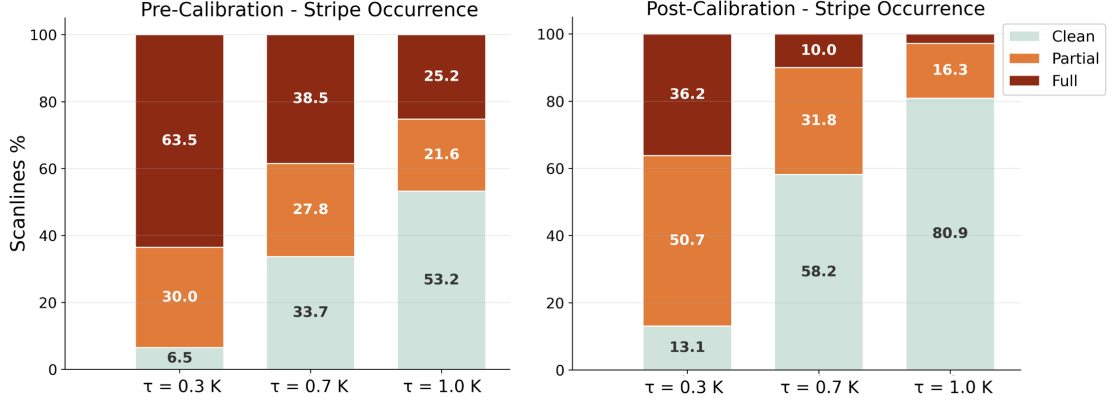


Figure 3.12: Stripe-class composition of the *Earth* view before (left) and after (right) two-point calibration, at the three error thresholds τ .

and $r = 0.91$ (*Hot-Cold*). The total noise shows almost identical correlations, confirming that the flicker component is the main driver of this shared structure.

Calibration study

To quantify the impact of calibration, the per-threshold fractions of *Clean*, *Partial* and *Full* scanlines in the *Earth* view are reported before and after calibration (Figure 3.12).

Two-point calibration is confirmed to suppress the flicker drift effectively at every threshold. At the 1 K threshold the clean scanline fraction rises from 53.2% to 80.9%; the effect is most pronounced on full stripes, which collapse from 25.2% to 2.8%, while partial stripes are reduced only mildly (from 21.6% to 16.3%). Calibration removes the coherent whole-scanline bias far more effectively than localized offsets, so that *partial* rather than *full* stripes become the dominant residual mode. The same pattern holds across thresholds.

Flagging study

The previous study established that calibration cannot fully suppress striping. A complementary operational tool would be a quality flag that warns downstream users when a given scanline is likely too corrupted to be trusted, allowing them to discard it rather than assimilate a contaminated observation.

The proxy is the moving average of $(\bar{n}_{\text{hot}} + \bar{n}_{\text{cold}})/2$ over 5 scanlines, where $\bar{n}$ is the mean of the total noise in each calibration window – the only quantity available operationally, since the flicker component cannot be isolated on orbit. Averaging over scanlines suppresses the thermal scatter and retains the slow flicker shared

Table 3.4: Pre-calibration operational warning matrix.

τ [K]	Flag	Flagged [%]	BT striped [%]	Recall [%]	Precision [%]	Lift
0.3	clean	30.1	87.6	71.8	96.1	1.03
	corrupt	69.9	96.1			
0.7	clean	62.1	51.3	51.9	90.9	1.37
	corrupt	37.9	90.9			
1.0	clean	78.8	34.7	41.6	92.0	1.97
	corrupt	21.2	92.0			

across the windows. An Earth view scanline is flagged as *corrupt* when the proxy exceeds τ , otherwise it is *clean*.

The flag is evaluated against the ground-truth stripe classification of the Earth view (BT striped), without distinguishing between full and partial stripes.

Three metrics quantify its effectiveness:

- *Precision*. Fraction of the scanlines flagged as corrupt that are truly striped; precision is related to the accuracy of the detection.
- *Recall*. Fraction of actual striped scanlines caught by the flag; high recall means few stripes are missed.
- *Lift*. Ratio of precision to overall striped fraction; it measures how much more likely a flagged scanline is to be striped than one picked at random.

A high precision alone is not sufficient to declare a flag useful: when the base rate is high, almost every scanline is striped and even a random flag appears precise. The flag is therefore informative only when recall exceeds the flagged fraction and lift exceeds unity.

Before calibration the usefulness of the proxy depends on the threshold (Table 3.4). At $\tau = 1.0$ K it fires on 21.2% of scanlines yet flagged scanlines are striped with 92% probability against a 46.8% base rate, a lift of 1.97. At $\tau = 0.7$ K the lift is 1.37, with a high precision of 91% against a 66.3% base rate. Its recall, however, is limited – 41.6% and 51.9% at the two thresholds – because the proxy sees only the flicker shared with the calibration windows and misses stripes driven by an excursion happened during the Earth view.

At the lowest threshold $\tau = 0.3$ K the proxy fires on 69.9% of scanlines and the lift collapses to 1.03: with a 93.5% base rate almost every scanline is striped, so flagging adds no information.

The proxy is therefore a useful but conservative warning: reliable when it fires, but

Table 3.5: Post-calibration operational warning matrix.

τ [K]	Flag	Flagged [%]	BT striped [%]	Recall [%]	Precision [%]	Lift
0.3	clean	30.1	87.5	69.8	86.8	1.00
	corrupt	69.9	86.8			
0.7	clean	62.1	41.3	38.7	42.8	1.02
	corrupt	37.9	42.8			
1.0	clean	78.8	19.2	20.7	18.7	0.98
	corrupt	21.2	18.7			

not a complete detector. This is visible in Figure 3.13: the flagged scanlines line up with the coherent flicker bands of the pre-calibration view, confirming that the flag captures the striping it is meant to detect.

After calibration, however, the proxy loses all predictive power (Table 3.5).

At every threshold the BT striped fraction is nearly identical between the two flag groups and the lift sits at 0.98–1.02, meaning the proxy catches stripes at almost exactly the rate it would by flagging scanlines at random; indeed, recall is comparable to the flagged fraction (20.7%, 38.7%, 69.8% against 21.2%, 37.9%, 69.9%). The reason is that calibration removes precisely the component common to all segments, leaving a residual dominated by flicker excursions that developed during the Earth observation - a contribution the calibration windows never see. The proxy therefore flags corrupted gains, not corrupted Earth views, and post-calibration these two sets are unrelated. This is reflected in Figure 3.14, where the flag removes scanlines essentially at random with respect to the residual striping.

Mitigation study

A final calibration improvement is tested: when a scanline is flagged as corrupt, the estimated gain is replaced by the linear interpolation from the two nearest clean neighbours, on the hypothesis that a smoother gain sequence might reduce the residual striping.

The comparison against standard calibration shows that gain interpolation does not help: at every threshold the fraction of striped scanlines increases slightly. At $\tau = 0.7$ K the striped fraction rises by 1.3 pp (from 41.8% to 43.1%); at $\tau = 0.3$ K and $\tau = 1.0$ K the increases are 0.5 pp and 1.1 pp respectively.

The result is consistent with the flagging study: the corrupted gain is at least contemporaneous with the Earth view it calibrates, while an interpolated one from adjacent scanlines carries a different flicker state. No gain-manipulation strategy can recover an excursion that the calibration windows never observed.

Overall, the four-stage analysis establishes the following operational conclusions:

- Flicker noise affects on average more than 20% of scanlines with a coherent error above 1 K.
- Two-point calibration is the most effective single intervention against striping, raising the clean scanline fraction on average from 50% to 80% at the 1 K threshold, collapsing full stripes from 25% to less than 3% and leaving partial stripes as the dominant residual mode.
- Calibration-window monitoring is a reliable quality flag before calibration – at $\tau = 1$ K a striped Earth view is flagged with a lift of almost 2 – but becomes uninformative on the calibrated image, where the residual striping is driven by flicker excursions happened in the Earth view alone.
- No gain-manipulation strategy improves on standard calibration. The residual striping must be addressed at the image level, which motivates the destriping methods of Chapter 5.

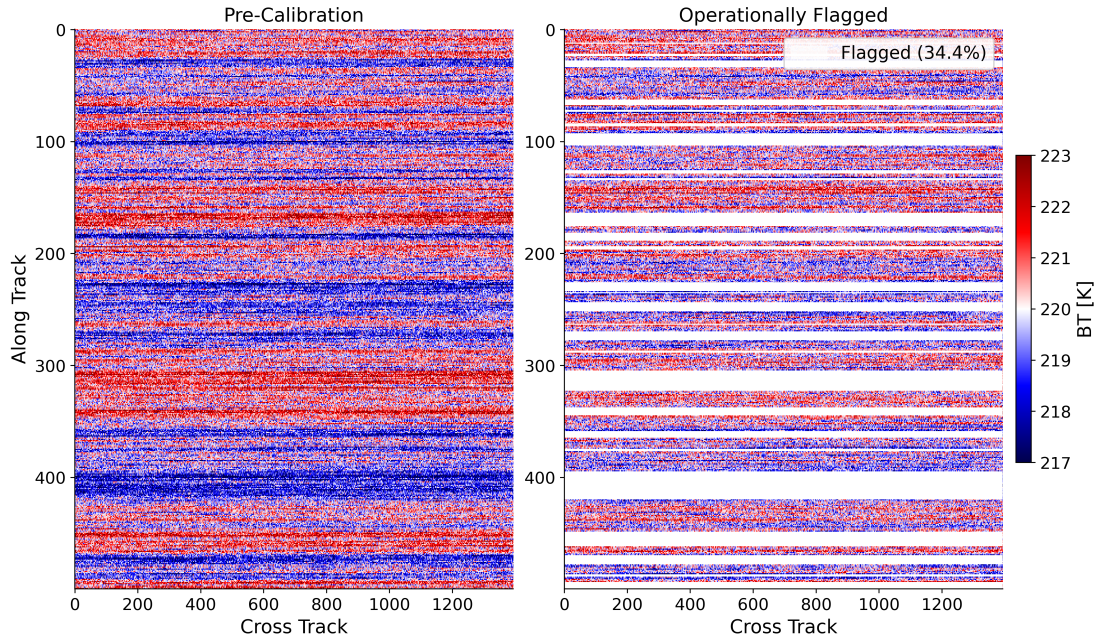


Figure 3.13: Pre-calibration Earth view (left) and the calibration-window flagging applied to it (right), at $\tau = 0.7$ K. Only the first 500 scanlines are displayed.

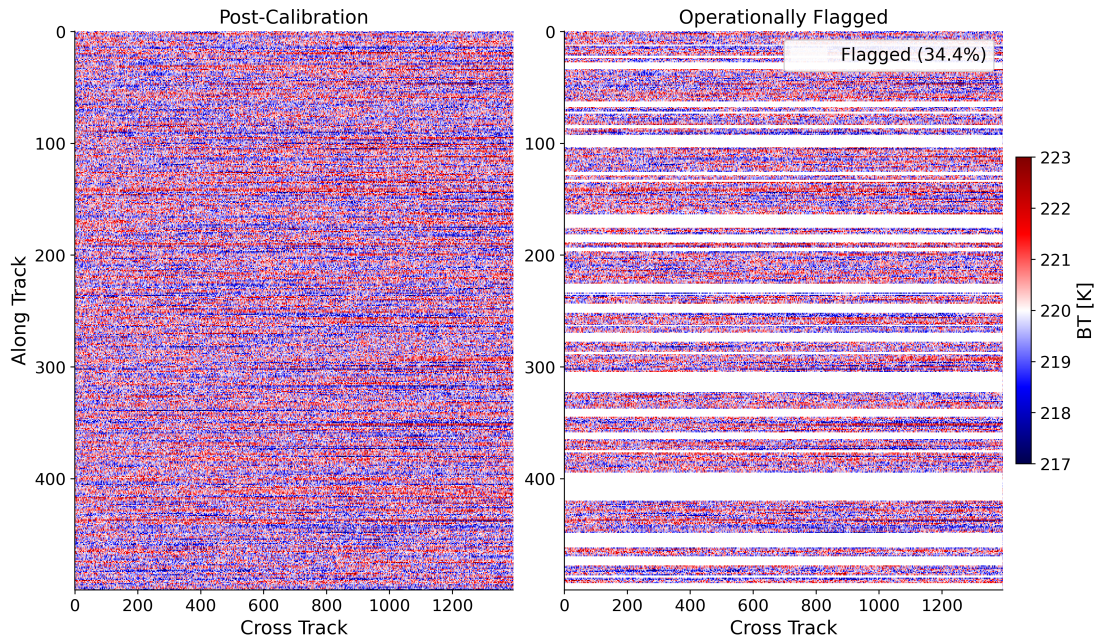


Figure 3.14: Post-calibration Earth view (left) and the calibration-window flagging applied to it (right), at $\tau = 0.7$ K. Only the first 500 scanlines are displayed.

Chapter 4

Synthetic Noise Generation

The noise model derived in Chapter 3 is now used to generate synthetic noise that reproduces the measured instrument behaviour. The goal is a generator whose noise, once injected into clean Earth observations, produces a realistic striped dataset for the challenge. Two approaches are developed and compared:

- *Parametric model.* Following the simulation method of Yang et al. [10], the noise is built from its two physical components.
- *Empirical model.* The noise is generated to reproduce the spectral structure extracted directly from the laboratory data.

Both operate in the frequency domain, which is the natural choice given that the reproducible quantity identified in Section 3.1.3 is the ensemble power spectral density rather than any time-domain realization.

4.1 Thermal+1/f model

The parametric approach constructs the noise as the sum of the two components identified in the model:

$$n(t) = n_{\text{thermal}}(t) + n_{1/f}(t). \quad (4.1)$$

The *thermal* component is additive white Gaussian noise, generated directly in the time domain as a zero-mean sequence scaled to the fitted standard deviation $\sigma_{\text{thermal}} = 0.564$ K.

The *flicker* component is generated in the frequency domain: a white Gaussian sequence is Fourier-transformed and each frequency bin is multiplied by $|f|^{\alpha/2}$, shaping the power spectral density to follow the measured power law with $\alpha = -1.30$; an inverse transform returns the time-domain signal, which is then scaled to the

Table 4.1: Calibration of the low-frequency cut-off f_c for the Thermal+1/ f generator.

f_c [Hz]	$R_{1/f}$ bias [%]	$P_{1/f}$ bias [%]	SI bias [%]
no cut-off	-1.31	-0.91	+9.81
0.0010	-1.52	-1.06	+6.07
0.0020	-1.71	-1.19	+2.88
0.0025	-1.89	-1.31	+0.83
0.0030	-2.02	-1.41	-0.76
0.0040	-2.06	-1.44	-1.78

fitted standard deviation $\sigma_{1/f} = 0.782$ K.

The values used to tune this model are estimated from the concatenated sequence, which gives more reliable estimates, and are indeed the same used to quantify the noise in Section 3.2.1. They differ only negligibly from those reported in Section 3.2.2, which are estimated on the individual acquisitions as a pre-processing step: $\sigma_{\text{thermal}} = 0.564$ K against 0.554 K and $\alpha = -1.30$ against -1.29 .

Validating the simulated against the observed noise shows that, although the PSD shape is reliably reproduced, the Allan deviation slopes and the Striping Index are not consistently matched. The next section introduces an additional step to resolve this discrepancy.

4.1.1 Cut-off frequency implementation

The source of the divergence is the low-frequency content: with no lower bound, the power-law spectrum diverges as the frequency approaches zero, injecting excessive low-frequency energy and producing an extremely slowly decaying ACF. To reproduce the physical behaviour a cut-off frequency f_c is introduced: the power-law shaping is applied only above f_c , and the flicker spectrum is flattened below it, bounding the low-frequency energy.

The cut-off is calibrated empirically by comparing the synthetic and observed striping metrics over a range of candidate values, repeating the comparison over 10 realizations to average out the realization-specific flicker excursions. Table 4.1 reports the result: without a cut-off the Striping Index is highly overestimated, and the bias decreases as f_c increases. The best trade-off across all metrics – minimizing the SI bias while keeping the Striping Percentage and Ratio within 2% – is given by $f_c = 0.0025$ Hz, which is therefore adopted for the parametric generator.

4.2 LUT model

The Look-Up Table (LUT) model was originally developed at ESA; the present work validated the existing implementation, corrected an improper bin normalisation in the construction of the LUTs, and carried out the window length selection.

Instead of exploiting the two-component structure, the LUT method reproduces the spectral structure directly from the laboratory measurements. The method has two stages:

1. *LUTs generation.* The concatenated sequence is divided into W non-overlapping windows of length N . Each window is zero-padded, Fourier-transformed, and the amplitude and phase of every frequency bin are extracted; then, two separate $(N - 1) \times W$ look-up tables are built – where $N - 1$ derives from dropping the DC bin – one storing the amplitudes and the other the phases. Under a Gaussian, stationary assumption the per-bin amplitudes and phases would follow known analytic distributions – Rayleigh and uniform, respectively; since the noise is only approximately Gaussian, the empirical amplitude and phase distributions are stored instead, preserving the actual departures observed in the data.
2. *Noise Synthesis.* A single window is generated by drawing, independently for each bin, a random amplitude and a random phase from the stored tables; by applying the inverse transform the time-domain segment is obtained. Sequences of arbitrary length are assembled by overlap-add of successive windows with a Hann window at a hop of $N/2$, ensuring amplitude continuity across boundaries. Unlike stage 1., here the Gaussian approximation is exploited: under it, amplitude and phase are statistically independent, which justifies drawing them separately.

The window length N is the only parameter of the model, and its selection is the subject of the next section. It is chosen by splitting the acquisitions into a training and a test set: the former builds the amplitude and phase LUTs for each candidate N , the latter serves as ground truth against which the synthetic noise is evaluated.

4.2.1 Window Length Selection

The window length N controls a fundamental trade-off. The frequency resolution of the table is $\Delta f = 1/(N \cdot T_{s,\text{lab}})$, so a longer window resolves lower frequencies and captures more of the $1/f$ structure that produces striping. At the same time, a longer length reduces the total number of windows W in the fixed-length dataset, degrading the reliability of each LUT entry. The objective is therefore to determine

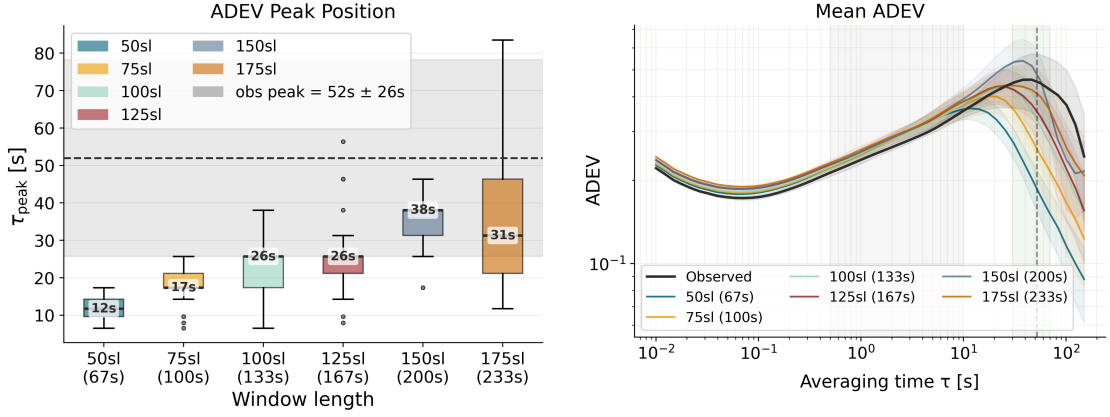


Figure 4.1: Window length validation of the LUT model through the Allan deviation. *Left:* ADEV peak position per N . *Right:* mean ADEV per N .

Table 4.2: Window length validation of the LUT model through the striping metrics.

Scanlines	N	$R_{1/f}$ bias [%]	$P_{1/f}$ bias [%]	SI bias [%]
50	66 650	−15.00	−11.40	−21.97
75	99 975	−7.41	−5.36	−15.12
100	133 300	−6.76	−4.79	−11.82
125	166 625	−6.58	−4.65	−7.63
150	199 950	+1.17	+1.02	−2.85
175	233 275	−1.85	−1.48	−6.67

the smallest N that still reproduces the spectral and spatial structure of the noise. A practical consequence of the long memory established in Chapter 3 is that the result depends on which flicker excursions happen to fall in the training and test sets. The selection is therefore not based on a single split but on a repeated random-subsampling cross-validation: 10 different train/test partitions are drawn, each assigning 8 of the 11 acquisitions to the training and the remaining 3 to the test, so that the chosen N is robust to the particular realizations present in any one split. Because whole acquisitions are assigned to either set, the continuity within each 200 s record is preserved. The LUT noise synthesis is itself stochastic, drawing random amplitudes and phases at each call, so 5 synthetic realizations are generated per partition and their metrics averaged.

After a first screening, values of N that are clearly too short or too long are discarded. The six retained candidates reproduce blocks of $\{50, 75, 100, 125, 150, 175\}$ scanlines, corresponding to $N \in \{66\,650, 99\,975, 133\,300, 166\,625, 199\,950, 233\,275\}$ samples

Table 4.3: Striping metrics of the two synthetic models against the observed noise.

Metric	Observed	LUT	Thermal+1/ f
$R_{1/f}$	1.3855	1.387 ± 0.059	1.359 ± 0.003
$P_{1/f}$	65.75	65.72 ± 1.90	64.87 ± 0.11
SI	2.1486	2.140 ± 0.122	2.159 ± 0.065

or equivalently $\{67, 100, 133, 167, 200, 233\}$ s. All retained candidates reproduce the PSD and the marginal distribution well, so the selection is driven by the Allan deviation and the three striping metrics (SI, $P_{1/f}$, $R_{1/f}$).

Windows shorter than 100 s can neither reproduce the Allan deviation up to the average peak nor match the striping metrics within a reasonable error. The striping patterns begin to look visually comparable to the observed ones starting from windows of 133 s, but the metrics converge only at 200 s. At this window length the synthetic Allan deviation peak coincides with the observed one (Figure 4.1) and the metric biases fall within a few percent (Table 4.2); increasing the length leaves too few training windows and indeed the metric biases increase again at 233 s.

The optimal value $N = 199\,950$ coincides with the duration of a single laboratory acquisition (200 s), which is both physically reasonable and expected. Once N is fixed, the training/test split is no longer needed and the final LUTs for synthetic noise generation are built on all the available data, for a total of $W = 11$ windows.

4.3 Models Comparison

With both generators calibrated, the parametric Thermal+1/ f and the empirical LUT models are compared against the observed noise, in terms of both their striping metrics and the temporal and spatial appearance of the noise.

The three striping metrics are reported in Table 4.3: both models reproduce the observed SI, $P_{1/f}$ and $R_{1/f}$ closely, with all biases below 2% over 10 realizations. The LUT method shows a larger spread (e.g. $R_{1/f} = 1.387 \pm 0.059$ against 1.359 ± 0.003 for the parametric model), an expected consequence of its empirical, realization-dependent nature. The distribution, PSD and Allan deviation are likewise reproduced by both models, as shown in Figure 4.2. The two methods differ, however, in the features that depend on the temporal continuity of the noise sequence, which the metrics above do not capture.

The first difference concerns the ACF, which is better reproduced by the Thermal+1/ f model (Figure 4.2). The reason is structural: the LUT method synthesizes each window independently, drawing amplitudes and phases separately for each bin with no constraint linking the phase of bin k in one window to its phase in the next.

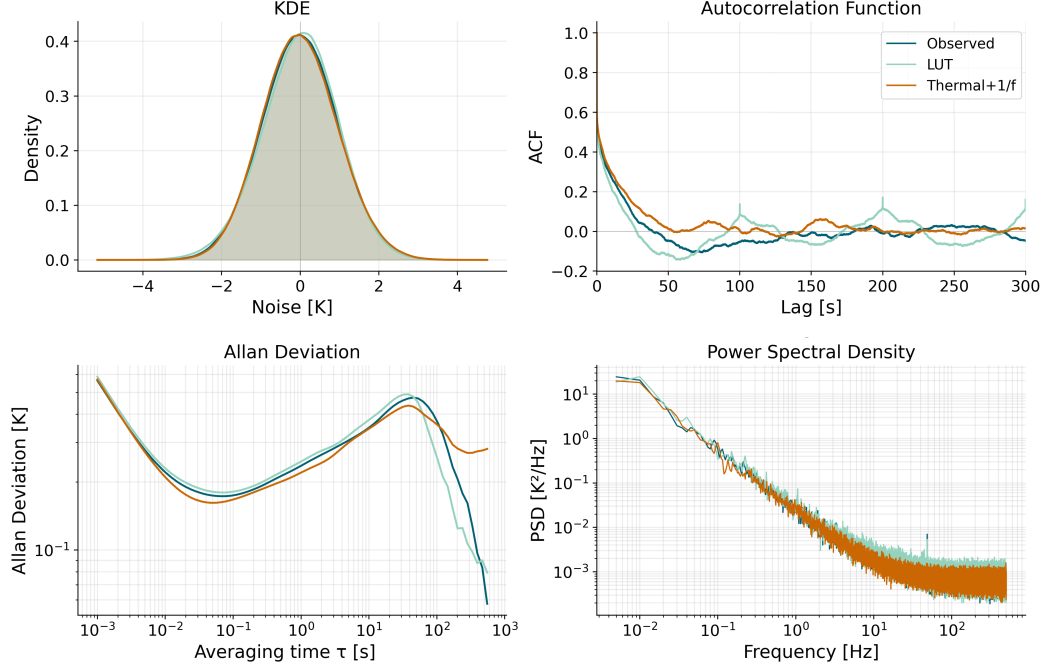


Figure 4.2: KDE, ACF, Allan Deviation and PSD of the observed, LUT and Thermal+ $1/f$ noise.

The Hann overlap-add enforces *amplitude* continuity across window boundaries but not *phase* continuity. This manifests in the ACF as oscillatory lobes with an amplitude larger than observed, because the phase randomization suppresses the interference patterns that would average out in a physically continuous process, and as spikes at the hop period, marking the window-pair boundary. Both artefacts are absent from the Thermal+ $1/f$ model, which generates a single continuous realization with no windowing structure.

The second difference concerns the appearance of the noise, both as a one-dimensional sequence and in its two-dimensional stripe pattern. In both views the LUT model resembles the observed data more closely than the parametric one: its drifts and fluctuations (Figure 4.3) and the resulting stripe pattern (Figure 4.4) are visually closer to the measurements, as expected given the empirical approach.

In summary, the two generators are statistically equivalent in their striping metrics and spectral content, and the choice between them is a trade-off: the LUT model reproduces the spatial appearance of the noise most faithfully, while the parametric Thermal+ $1/f$ model reproduces the autocorrelation structure and offers full control over the noise composition.

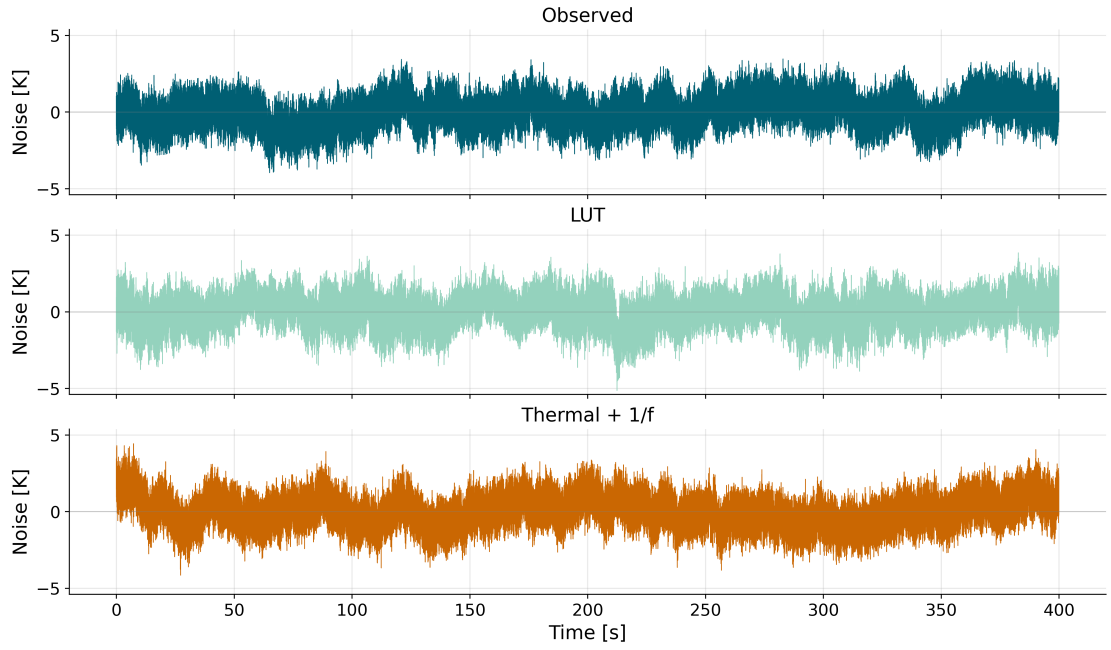


Figure 4.3: Time-series comparison of a noise segment for the observed data and the LUT and Thermal+1/ f models.

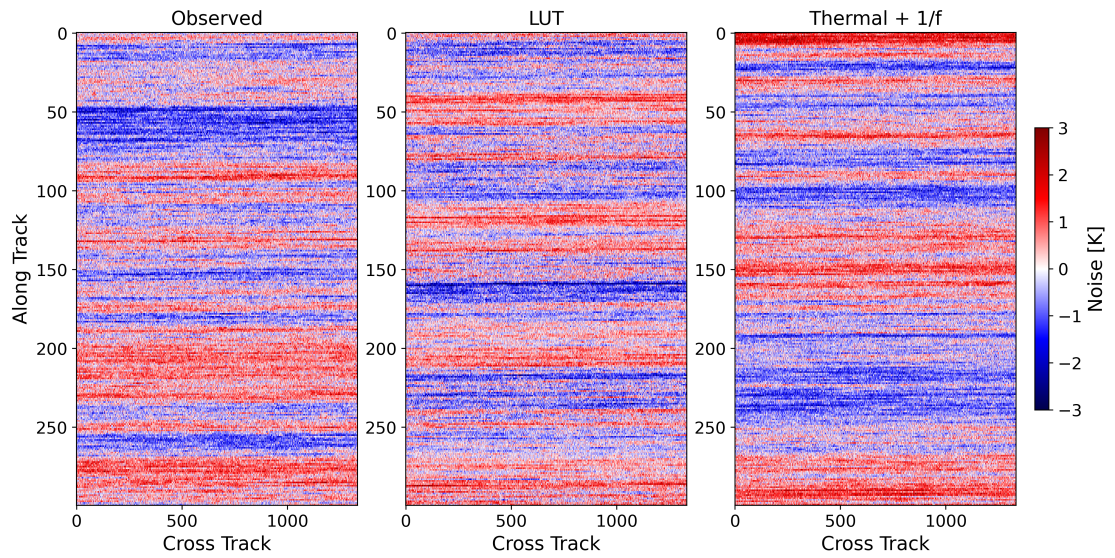


Figure 4.4: Corresponding swath of Figure 4.3 obtained by folding each series into the scan geometry.

Chapter 5

Noise Mitigation

This chapter addresses the mitigation of the residual striping that, as established in Chapter 3, calibration alone cannot remove.

The methods selected in Chapter 2 – the Wiener filter, PCA/EEMD and BM3D – are described and implemented in turn, together with a hybrid method combining PCA/EEMD and BM3D.

Their evaluation on simulated Earth scenes and their use in the preparation of the challenge are addressed in Chapter 6.

Notation

The following notation is used throughout the chapter:

- L : the along-track dimension, i.e. the number of scanlines.
- M : the cross-track dimension, i.e. the number of field of views across the swath, fixed by the instrument.
- $X, N, Y \in \mathbb{R}^{L \times M}$: the matrices representing the clean swath, the noise and the noisy swath respectively, related by the additive model $Y = X + N$ (Equation 2.1).
- $\hat{X} \in \mathbb{R}^{L \times M}$: the denoised estimate returned by a destriping method.
- $x, n, y, \hat{x} \in \mathbb{R}^{LM}$: corresponding signals, used by methods that operate in one dimension.

5.1 Wiener Filter

The Wiener filter is the linear minimum-mean-square-error (MMSE) estimator of a target process from the measurements of another process jointly wide-sense stationary with it [26].

The estimate is obtained as the output of an LTI (linear and time-invariant) system applied to the measurement, $\hat{x} = h * y$, where the impulse response h is the one minimising the mean-square error

$$\min_{h[\cdot]} \mathbb{E}[e^2[n]], \quad e[n] = \hat{x}[n] - x[n].$$

Setting the derivative with respect to the filter coefficients to zero yields the orthogonality condition, which in the frequency domain gives the optimal transfer function $H = S_{xy}/S_y$, the ratio of the signal–measurement cross-spectrum to the measurement spectrum. When the signal and the noise are uncorrelated, this reduces to $H = S_x/(S_x + S_n)$.

For the destriping problem we are in the additive setting of Equation 2.1, with uncorrelated clean signal X and noise N . Indeed, the source of the noise lies in the receiver electronics, independently of the scene content; the only coupling between the two is the dependence of the noise amplitude on scene temperature established in Section 3.2.2, which is handled separately as described below.

The Wiener filter therefore takes the form

$$H(f) = \frac{S_x(f)}{S_x(f) + S_n(f)}, \quad (5.1)$$

where $S_x(f)$ and $S_n(f)$ are the power spectral densities of the clean signal and the noise. The filter passes the frequencies where the signal dominates ($H \rightarrow 1$) and suppresses those where the noise dominates ($H \rightarrow 0$).

As a one-dimensional filter, it is applied at the acquisition level: the swath is processed by rows and folded back into its two-dimensional form after filtering.

Implementing the filter requires addressing three aspects:

- *Unknown signal spectrum.* The true underlying Earth signal is not known, so its PSD $S_x(f)$ is estimated from the observation by spectral subtraction

$$S_x(f) = \max\{S_y(f) - S_n(f), 0\}, \quad (5.2)$$

where the clipping at zero guarantees a non-negative spectrum. While the noise PSD $S_n(f)$ is estimated once over the entire dataset and considered known, the observation PSD $S_y(f)$ is estimated within each window from a single realization, and is therefore smoothed over a small frequency band to reduce its variance before the subtraction.

- *Local stationarity and windowing.* A single global transform would assume the signal and noise to be stationary over the whole swath, a requirement clearly violated as the scene varies and the noise amplitude with it. The sequence is therefore processed by overlap-add (OLA) over windows short enough that the signal statistics are approximately stationary within each: every window is Hann-tapered, transformed to the frequency-domain, filtered, and recombined in the time-domain at a 50% hop. The window length N_{win} involves a trade-off: on one side the stationarity assumptions limits the extent of the window; on the other, the frequency resolution is $\Delta f = 1/(N_{\text{win}} \cdot T_{\text{op}})$, so a longer window resolves the low-frequency band where the striping concentrates. The choice of N_{win} is addressed in Section 6.2.
- *Amplitude scaling.* The noise PSD is estimated once at the laboratory operating point, fixing the shape of the noise, but the amplitude scales with the scene temperature (Section 3.2.2). Hence, within each window the reference PSD is rescaled to the local scene temperature through Equation 3.6

$$S_n^{\text{win}}(f) = S_n(f) \cdot \left(\frac{T_{\text{n,rec}} + \bar{T}_{\text{win}}}{T_{\text{sys,lab}}} \right)^2, \quad (5.3)$$

where $\bar{T}_{\text{win}}$ is the mean brightness temperature of the window.

Algorithm 1 Wiener Filter (window overlap-add)

Require: noisy image $Y \in \mathbb{R}^{L \times M}$; reference noise PSD S_n ; window length N_{win}

- 1: $y \leftarrow \text{ravel}(Y)$
- 2: remove global linear trend from y ; edge-pad by $N_{\text{win}}/2$
- 3: $w \leftarrow$ Hann window of length N_{win} ; hop $\leftarrow N_{\text{win}}/2$
- 4: **for** each window y_k at hop spacing **do**
- 5: $\hat{S}_y \leftarrow \text{Welch}(y_k - \bar{y}_k)$; smooth $\hat{S}_y$
- 6: $S_n^{\text{win}} \leftarrow S_n \cdot \left((T_{\text{n,rec}} + \bar{T}_k) / T_{\text{sys,lab}} \right)^2$ $\triangleright$ scale to local scene temp
- 7: $\hat{S}_x \leftarrow \max\{\hat{S}_y - S_n^{\text{win}}, 0\}$
- 8: $H \leftarrow \hat{S}_x / (\hat{S}_x + S_n^{\text{win}})$
- 9: $\hat{y}_k \leftarrow \text{IFFT}(H \cdot \text{FFT}(w \cdot (y_k - \bar{y}_k)))$
- 10: accumulate $w \cdot \hat{y}_k$ into output; accumulate w^2 into the normaliser
- 11: **end for**
- 12: $\hat{x} \leftarrow \text{output}/\text{normaliser}$; remove padding; restore global trend
- 13: $\hat{X} \leftarrow \text{reshape } \hat{x} \text{ to } L \times M$
- 14: **return** $\hat{X} \in \mathbb{R}^{L \times M}$

5.2 PCA/EEMD

As outlined in Chapter 2, the combination of Principal Component Analysis (PCA) and Ensemble Empirical Mode Decomposition (EEMD) is the established destriping benchmark for microwave radiometers.

The methods are briefly explained below, since the analysis performed by Qin et al. [14] on the ATMS and by Dong and Zou [16] on the GMI is applied to MWI-13 V striping noise.

Principal Component Analysis

Principal Component Analysis is generally known as a dimensionality reduction technique, which retains the variables explaining the most variance.

In the case of destriping, the leading components found by PCA capture the highest along-track variance, which corresponds to the striping pattern. Rather than retaining these components, PCA is therefore used to locate the noise, so that it can be isolated and removed.

Let $A = Y^\top \in \mathbb{R}^{M \times L}$ be the matrix of brightness temperature observations of the selected portion of swath, where columns represent scanlines.

PCA proceeds by diagonalising the scatter matrix $S = AA^\top \in \mathbb{R}^{M \times M}$ to find the directions of maximum variance: its eigenvectors, here called the *Principal Component (PC) modes* (i.e. loadings), are ordered by decreasing eigenvalue and stored as the columns of $E = (e_1, \dots, e_M) \in \mathbb{R}^{M \times M}$. The i -th mode $e_i \in \mathbb{R}^M$ describes the i -th cross-track structure of the brightness temperatures.

The observations are then projected onto the modes, yielding the *PC coefficients* (i.e. scores), stored as the rows of $U = (u_1, \dots, u_M)^\top \in \mathbb{R}^{M \times L}$. The i -th row $u_i \in \mathbb{R}^L$ is the weight of mode i across the swath, describing its along-track variation.

Combining modes and coefficients, the swath matrix A can be written as the sum of rank-one *PCA components*, each the outer product of a cross-track mode (a column of E) and its along-track coefficient (the corresponding row of U)

$$A = E U = \sum_{i=1}^M e_i u_i, \quad (5.4)$$

so that, by removing the contribution of the striping noise from the leading coefficients of U , a destriped estimate of A can be reconstructed.

As shown in Figure 5.1, the leading component indeed captures almost all of the striping: the explained variance is above 99.99% for PC mode 1, dropping to the order of $10^{-5}\%$ for modes 2 and 3.

The number of PCA components from which to extract the noise, hereafter N_{PC} , is therefore the first parameter. Qin et al. [14] only consider the first component for

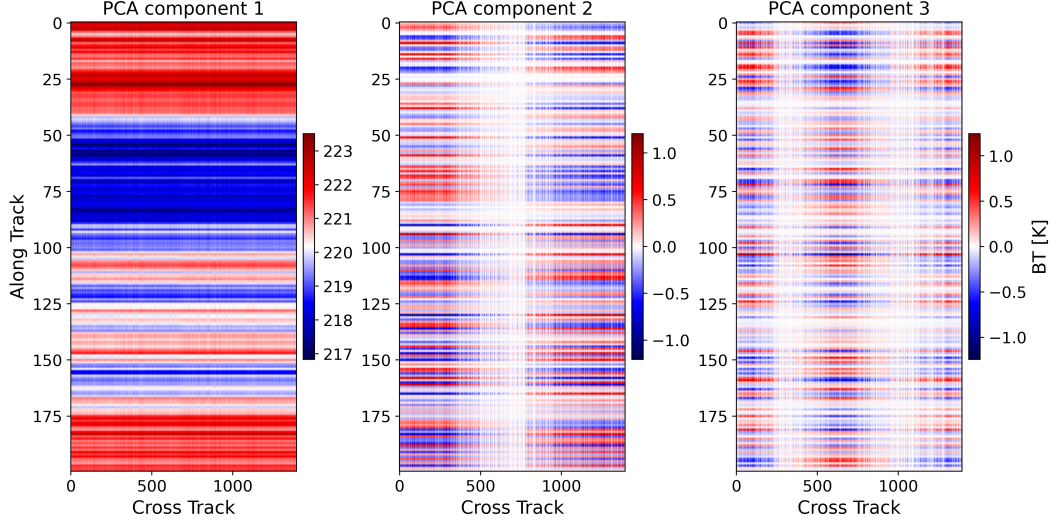


Figure 5.1: First, second and third PCA components of ground-test MWI-13 V noise injected on an Earth scene at constant temperature (220 K).

the ATMS striping noise, discarding the second and third due to their negligible explained variance; Dong and Zou [16], instead, retain all three for the GMI.

Ensemble Empirical Mode Decomposition

Ensemble Empirical Mode Decomposition (EEMD) is a noise-assisted data analysis (NADA) method, developed by Wu and Huang [27] as an improvement of the Empirical Mode Decomposition (EMD). EMD adaptively decomposes a one-dimensional signal into a set of oscillatory modes, the *Intrinsic Mode Functions* (IMFs), ordered from highest to lowest frequency, plus a residual trend; the IMFs are extracted from the extrema of the signal, with no prior choice of basis functions. EEMD produces an ensemble of perturbed signals by adding a different white noise realization to the original series many times and averaging the resulting IMFs: this suppresses the mode-mixing that affects EMD, while the added noise cancels out in the average.

In the destriping case, EEMD is applied to a leading PC coefficient u_i , decomposing it as

$$u_i = \sum_{m=1}^n C_m^{(i)} + R_n^{(i)}, \quad (5.5)$$

where the IMFs $C_m^{(i)}$ are ordered from the highest frequency ($m = 1$) to the lowest, n is the total number of extracted modes, and $R_n^{(i)}$ is the residual trend.

The striping noise is shown to be captured by the first, high-frequency IMFs; by

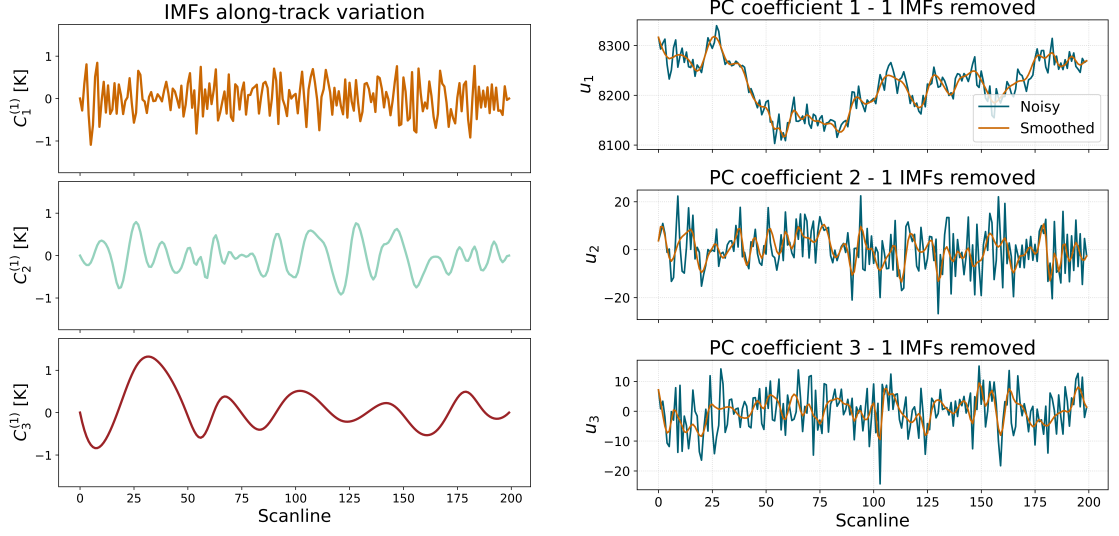


Figure 5.2: *Left:* along-track variation of the first three high-frequency IMFs $C_m^{(1)}$ extracted from the first PC coefficient u_1 , at nadir ($e_{697,1}C_m^{(1)}$, $m = 1,2,3$). *Right:* the first three PC coefficients before (Noisy) and after (Smoothed) the removal of their respective IMFs.

subtracting such modes from the coefficient, a smoothed estimate is obtained

$$\hat{u}_i = u_i - \sum_{m=1}^{N_{IMF}^{(i)}} C_m^{(i)} \quad (5.6)$$

and used to reconstruct a denoised swath through Equation 5.4.

This behaviour is illustrated in Figure 5.2 in the case of MWI-13 V noise. The along-track variation of the IMFs of the first PC coefficient (*left*) confirms the expected ordering: the first IMF is the highest-frequency mode, and the frequency content decreases progressively in the second and third modes. Subtracting these high-frequency IMFs from the PC coefficients (*right*) smooths their along-track variation while leaving the underlying trend intact.

The number of IMFs subtracted from each PC coefficient u_i , $N_{IMF}^{(i)}$, is the second parameter of the method, and the two reference works determine it differently. Qin et al. [14] inspect the along-track variation of the first PCA component as IMFs are progressively removed; they set $N_{IMF}^{(1)} = 3$, as it leads to a smooth profile. Dong and Zou [16], instead, rely on a spectral analysis, examining the peaks of the PSD of the first four IMFs of the first three PC coefficients; they determine striping noise exists in the first two IMFs, so they set $N_{IMF}^{(i)} = 2$, $i = 1,2,3$. The two criteria are reproduced for MWI-13 V noise in Figure 5.3. By adapting the

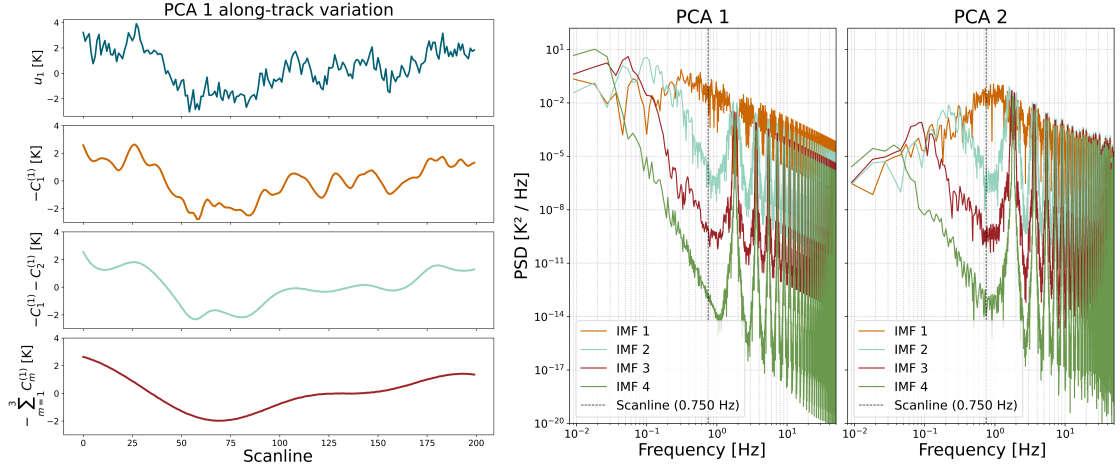


Figure 5.3: *Left* (ATMS approach): along-track variation of the first PCA component at nadir, progressively removing the first three IMFs. *Right* (GMI approach): PSD of the first four IMFs at nadir (PCA 3 is omitted, as it is similar to PCA 2).

same reasoning, fewer IMFs are needed: for the approach used on the ATMS (*left*), removing the third IMF oversmooths the profile, so only the first two are subtracted ($N_{IMF}^{(1)} = 2$); for the approach used on the GMI (*right*), the PSDs of IMFs 2, 3 and 4 peak at low frequency, indicating those modes carry signal rather than striping, so only the first is removed from each coefficient ($N_{IMF}^{(i)} = 1$, $i = 1, 2, 3$).

Algorithm 2 PCA/EEMD

Require: noisy image $Y \in \mathbb{R}^{L \times M}$; PC modes N_{PC} ;

IMFs to remove per mode $\{N_{IMF}^{(i)}\}$

- 1: $A \leftarrow Y^\top$ $\triangleright$ columns = scanlines
 - 2: $E, U \leftarrow \text{PCA}(A)$ $\triangleright S = AA^\top = E\Lambda E^\top$; $U = E^\top A$
 - 3: $\hat{U} \leftarrow U$
 - 4: **for** $i = 1$ to N_{PC} **do**
 - 5: $\{C_m^{(i)}\} \leftarrow \text{EEMD}(u_i)$
 - 6: $\hat{u}_i \leftarrow u_i - \sum_{m=1}^{N_{IMF}^{(i)}} C_m^{(i)}$ $\triangleright$ remove the $N_{IMF}^{(i)}$ highest-frequency IMFs
 - 7: **end for**
 - 8: $\hat{X} \leftarrow (E\hat{U})^\top$
 - 9: **return** $\hat{X} \in \mathbb{R}^{L \times M}$
-

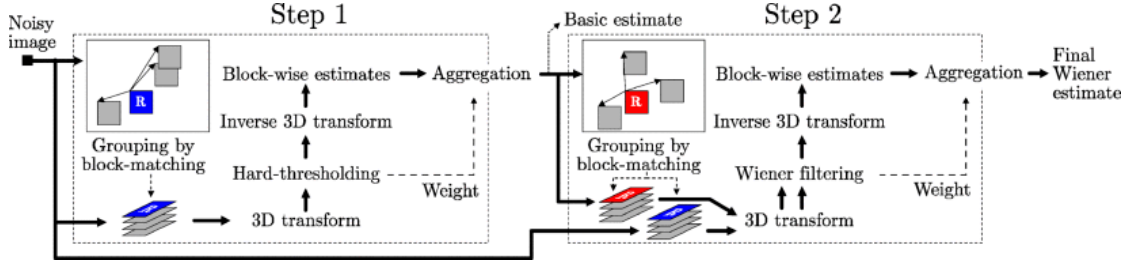


Figure 5.4: Flowchart of the BM3D image denoising algorithm. Source: adapted from [19].

5.3 BM3D

Block-Matching and 3-D filtering (BM3D) is an image denoising algorithm introduced by Dabov et al. [19]. The key concept it builds on is *similarity*: a given patch of a scene usually resembles many other patches found elsewhere in the image. While true signal and noise can be almost indistinguishable locally, a collection of similar patches shares a common underlying structure while carrying different noise realizations. Filtering the patches jointly therefore attenuates the noise while preserving the features of the signal.

Formally, the method follows three operations:

- *Grouping*. Each patch of the image is in turn taken as a *reference*, and the patches most similar to it are collected and stacked into a 3-D group. This is performed by *block-matching* (BM), with pairwise similarity measured by ℓ^2 -distance within a fixed threshold; since matching does not partition the image, the groups are not necessarily disjoint.
- *Collaborative filtering*. Given a group, each 2-D fragment collaborates in the filtering of all the others. Specifically, the denoised estimate is obtained by shrinkage in a sparsifying 3-D transform domain; sparsity is achieved by exploiting both the intra-fragment and inter-fragment correlation.
- *Aggregation*. As BM does not impose partitioning, a patch can – and generally will – belong to several groups and is therefore estimated more than once. The final estimate is obtained by weighted averaging, with noisier estimates given smaller weight.

The algorithm applies these operations in two steps (Figure 5.4):

1. *Basic estimate*. For each 3-D group G_{basic}^i formed from the noisy image Y , the

3-D coefficients are shrunk by hard-thresholding

$$\hat{G}_{\text{basic}}^i = \mathcal{T}_{3\text{D}}^{-1}\left(\Upsilon\left(\mathcal{T}_{3\text{D}}(G_{\text{basic}}^i)\right)\right), \quad (5.7)$$

where $\mathcal{T}_{3\text{D}}$ is the 3-D transform and Υ sets to zero every coefficient below a threshold proportional to the noise level. Aggregating the filtered groups $\hat{G}_{\text{basic}}^i$ yields the basic estimate $\hat{X}^{\text{basic}}$.

2. *Final estimate.* Assuming step 1. significantly attenuated the noise, BM is repeated on $\hat{X}^{\text{basic}}$ and the obtained locations are used to form groups on both the basic estimate and the original noisy image, $G_{\text{final,est}}^i$ and $G_{\text{final,noisy}}^i$. Using the energy spectrum of the basic estimate as the true signal spectrum, the hard-thresholding is replaced by an empirical Wiener filtering on the noisy group $G_{\text{final,noisy}}^i$

$$W = \frac{\left|\mathcal{T}_{3\text{D}}(G_{\text{final,est}}^i)\right|^2}{\left|\mathcal{T}_{3\text{D}}(G_{\text{final,est}}^i)\right|^2 + \sigma^2}, \quad (5.8)$$

where σ^2 is the noise variance. The inverse transform and a second aggregation yield the final estimate $\hat{X}$.

The original formulation assumes the image to be corrupted by i.i.d. zero-mean Gaussian noise, so that a single variance σ^2 governs the shrinkage in Equations 5.7 and 5.8. This assumption fails in many realistic settings, from medical to satellite imaging. Mäkinen et al. [20] relaxed it by extending the method to stationary additive correlated noise with known PSD (non-blind denoising): the scalar variance is replaced by the noise variance computed exactly in the transform domain from the 2-D noise PSD, so that the threshold and the Wiener coefficients adapt to the spectral shape of the noise. This is precisely the case of the MWI, whose noise is correlated and well characterised by a known PSD, as established in Chapter 3.

Similarly to the Wiener filter, a few aspects must be addressed to apply BM3D to the destriping problem:

- *Noise PSD.* The implementation of the algorithm relies on the **bm3d** package [20], which only requires the two-dimensional noise PSD as input parameter. The one-dimensional ground-test noise is shaped into its two-dimensional form by reproducing the acquisition geometry (scanning and calibration, assuming a constant scene), in order to match how it enters a calibrated swath. The obtained noise image is split into (H, W) blocks whose periodograms are averaged into the reference PSD $S_n \in \mathbb{R}^{H \times W}$.
- *Amplitude scaling.* The PSD is estimated once, fixing the shape of the noise, but the amplitude scales with the scene temperature (Section 3.2.2). Hence,

the swath is processed in blocks and within each the reference PSD is rescaled through Equation 3.6

$$S_n^{\text{blk}} = S_n \cdot \left(\frac{T_{\text{n,rec}} + \tilde{T}_{\text{blk}}}{T_{\text{sys,lab}}} \right)^2, \quad (5.9)$$

where $\tilde{T}_{\text{blk}}$ is the median brightness temperature of the block.

- *Blocking and overlap-add.* The blocks are recombined by two-dimensional overlap-add (OLA) with a separable Hann window, in both the along-track and cross-track directions. The block size (H, W) involves the usual trade-off: small blocks keep the amplitude scaling precise, while large blocks give BM3D enough patches for effective grouping.

Algorithm 3 Reference PSD estimation

Require: 1-D ground-test noise n ; block size (H, W)

- 1: build noise scan sequence over a constant scene $T_{\text{ref}} = T_{\text{sys,lab}} - T_{\text{n,rec}}$
 - 2: calibrate the sequence $\rightarrow$ noise image $N \triangleright$ matches noise in a calibrated swath
 - 3: $S_n \leftarrow$ average of $|\text{FFT2}(N_{ij} - \bar{N}_{ij})|^2$ over all blocks of N
 - 4: **return** $S_n \in \mathbb{R}^{H \times W}$
-

Algorithm 4 BM3D (block overlap-add)

Require: noisy image $Y \in \mathbb{R}^{L \times M}$; reference noise PSD $S_n \in \mathbb{R}^{H \times W}$;

block size (H, W)

- 1: $w \leftarrow$ separable 2-D Hann window of size (H, W) ; $\text{hop} \leftarrow (H/2, W/2)$
 - 2: **for** each block Y_k at hop spacing **do**
 - 3: $\tilde{T}_k \leftarrow \text{median}(Y_k)$; $Z_k \leftarrow Y_k - \tilde{T}_k$ $\triangleright$ center the block
 - 4: $S_n^{\text{blk}} \leftarrow S_n \cdot \left((T_{\text{n,rec}} + \tilde{T}_k) / T_{\text{sys,lab}} \right)^2$ $\triangleright$ scale PSD to local scene temp
 - 5: $\hat{Z}_k^{\text{basic}} \leftarrow$ hard-thresholding of Z_k $\triangleright$ step 1, Eq. 5.7
 - 6: $\hat{\hat{Z}}_k \leftarrow$ Wiener filtering of Z_k guided by $\hat{Z}_k^{\text{basic}}$ $\triangleright$ step 2, Eq. 5.8
 - 7: $\hat{X}_k \leftarrow \hat{\hat{Z}}_k + \tilde{T}_k$ $\triangleright$ restore level
 - 8: accumulate $w \cdot \hat{X}_k$ into output; accumulate w into the normaliser
 - 9: **end for**
 - 10: $\hat{X} \leftarrow$ output/normaliser
 - 11: **return** $\hat{X} \in \mathbb{R}^{L \times M}$
-

5.4 Hybrid Method: PCA/EEMD+BM3D

As analysed in Chapter 6, PCA/EEMD and BM3D fail on complementary parts of the problem: the first removes striping effectively but can also affect scene structure; the second works well as a general denoiser but oversmooths the scene while only partly suppressing the banding. A hybrid method is therefore proposed, as combining the two allows each to compensate for the other’s weakness.

The two methods are applied in sequence: PCA/EEMD acts first, on the full image, removing the coherent along-track striping; it is run in its gentlest configuration ($N_{PC} = 1$, $N_{IMF}^{(1)} = 1$), so that only the dominant stripe mode is removed. BM3D then refines the result, but with a weaker hard-threshold and a higher Wiener regularization, trusting the first-stage estimate more ($\lambda_{3D} = 0.8$, $\mu_2 = 1.3$ against the default 2.7 and 1.0). Each stage is therefore deliberately mild: PCA/EEMD does not have to clear the striping completely, since BM3D removes what remains, and BM3D does not have to smooth aggressively, since most of the striping is already gone.

Algorithm 5 PCA/EEMD+BM3D (Hybrid)

Require: noisy image $Y \in \mathbb{R}^{L \times M}$;

PC modes N_{PC} ; IMFs to remove per mode $\{N_{IMF}^{(i)}\}$

reference noise PSD $S_n \in \mathbb{R}^{H \times W}$; block size (H, W) ; λ_{3D}, μ_2

- 1: $Y_{\text{pca}} \leftarrow \text{PCA/EEMD}(Y; N_{PC}, \{N_{IMF}^{(i)}\})$ $\triangleright$ Stage 1: remove coherent striping
 - 2: $\hat{X} \leftarrow \text{BM3D}(Y_{\text{pca}}; S_n, (H, W), \lambda_{3D}, \mu_2)$ $\triangleright$ Stage 2: refine residual
 - 3: **return** $\hat{X} \in \mathbb{R}^{L \times M}$
-

Chapter 6

Challenge Preparation and Results

The four methods implemented in Chapter 5 are now applied to simulated MWI scenes, with a twofold aim: establish which performs best in an operational setting and prepare the scientific challenge.

The chapter therefore builds the challenge dataset, defines the evaluation metrics and the scene-adaptive score derived from them, and evaluates the methods to fix a baseline. Finally, the baseline’s residual failure is analysed and a candidate refinement proposed.

6.1 Challenge Dataset

The challenge dataset is built by injecting the validated synthetic noise of Chapter 4 onto clean MWI Level-1B scenes¹. The knowledge of both X and N is what makes the open competition possible.

6.1.1 MWI L1B Test Data

The effectiveness of a destriping algorithm lies in its ability to remove striping without distorting the underlying weather signal: a removed or altered geophysical structure degrades NWP just as a striping artefact misread as a real feature would. Therefore, the dataset should be representative of the real operational setting the algorithms are developed for.

¹Satellite data products are organised by processing level. Level-0 (L0) is the raw, uncalibrated instrument output; Level-1B (L1B) is the calibrated and geolocated measurement; Level-2 (L2) is the retrieved geophysical quantity derived from L1B, such as water vapour or precipitation.

As the first MetOp-SG B satellite (B1) is scheduled for launch only in late 2026, no real MWI observations were available during this work. The starting point for the dataset is therefore the MWI Level-1B (L1B) Test Data distributed by EUMETSAT²: a set of simulated orbits generated with the MWI L1B In-house Prototype Processor (IPP) developed at EUMETSAT, in which radiance values are associated with the geolocation and calibration processing prescribed by the EPS-SG MWI L1B Product Format Specification. The data are based on the MetOp A orbits 4655 and 4656 of 12 September 2007 and orbit 6985 of 23 February 2008, selected to span the full range of MWI observing conditions, from the illuminated African desert to the cold Antarctic surfaces [28].

Each orbit is stored in a single NetCDF file. The product reports *radiance* R (provided in $\text{mW m}^{-2} \text{sr}^{-1} (\text{cm}^{-1})^{-1}$) rather than brightness temperature T_B (in K), which is recovered through the Planck inversion

$$T_B(\nu_{ci}, R) = \frac{c_2 \nu_{ci}}{\ln\left(1 + \frac{c_1 \nu_{ci}^3}{R}\right)} \cdot A_i + B_i, \quad (6.1)$$

where $c_1 = 1.19 \times 10^{-5} \text{ mW}/(\text{sr m}^2 \text{ cm}^{-4})$ and $c_2 = 1.44 \text{ K cm}$ are the radiation constants and A_i , B_i , ν_{ci} are per-channel coefficients, equal to 1.0, 0.0 K and 5.52 cm^{-1} for MWI-13 V. The warm-load reference T_w used in the calibration of Section 3.2.2 and in the noise injection below is not assumed but read directly from the product: the on-board calibration target (OBCT) temperature is stable across all three orbits, with a mean of 282.05 K.

The three orbits contain 4458, 4447 and 4450 scanlines respectively, each of 1392 Earth view samples, together covering roughly five hours of acquisition. The 1392 Earth samples per scanline differ from the 1394 assumed in Chapter 1; the value reported in the Product Format Specification is still marked to-be-confirmed and the two-sample mismatch is reconciled at injection as described in Section 6.1.2.

The along-track and cross-track sampling are markedly anisotropic. Computing the ground-track speed directly from the navigation data gives $v_{\text{ground}} \approx 6.66 \text{ km/s}$ which, over the scanline duration of $\approx 1.33 \text{ s}$, corresponds to an along-track pixel spacing of $\approx 8.88 \text{ km}$; the cross-track spacing, by contrast, is only $\approx 1.22 \text{ km}$ for the 1700 km swath, so the along-track pixel is about seven times larger than the

²Available through the EUMETSAT User Portal, <https://user.eumetsat.int/>.

cross-track one. This sampling spacing should not be confused with the instantaneous field of view, which at 165 GHz is roughly 10 km, meaning that adjacent cross-track samples overlap heavily.

This anisotropy, together with the requirement that each image in the dataset contains enough scanlines for the striping to be characterisable, drives the choice of the patch size described next.

6.1.2 Dataset Generation

Noise is injected onto each clean orbit by reproducing the full acquisition chain, ensuring that it acquires the correct striping structure.

The empirical LUT generator (Section 4.2) is used as the noise source, as it reproduces the spatial appearance of the noise most faithfully. For each orbit the procedure is:

1. *Load and convert data.* The orbit radiance is read and converted to brightness temperature through Equation 6.1, giving the clean orbit in K.
2. *Build the LUTs.* The amplitude and phase look-up tables are built once, from all available (22 °C, *Meas*) laboratory data.
3. *Synthesise noise.* A continuous noise sequence long enough to cover the whole orbit is generated at the laboratory sampling rate and resampled to the operational rate $T_{s,op} = 0.394$ ms (Section 3.2.2).
4. *Fold and calibrate.* The sequence is folded into the scan geometry, each *Cold*, *Hot* and *Earth* segment is rescaled to its physical temperature through Equation 3.6, and the two-point calibration is applied producing the calibrated noisy brightness temperature orbit.
5. *Patch.* The orbit is partitioned into fixed-size images, as described below.

The synthetic scan geometry produces 1394 earth samples per scanline against the 1392 of the L1B product; the mismatch is reconciled by symmetric edge-padding of the clean orbit before injection and cropping one sample from each cross-track edge of the result, so that the noise-injected orbit matches the noise-free one exactly. To guarantee independent noise realizations a distinct random seed is used for each orbit.

To obtain a dataset from the full orbits, each is divided into non-overlapping rectangular patches. Three considerations constrain the patch geometry:

1. *Cross-track extent.* The 1392 cross-track samples are few and fixed, so the patch width is chosen from the divisors of 1392, namely $\{87, 116, 174, 232, 348, 464, 696\}$, to partition the swath without discarding any sample.
2. *Along-track extent.* Each patch must span enough scanlines for the striping to be characterisable, since striping is a per-scanline effect and the destriping algorithms need sufficient along-track extent to act on (i.e. for PCA, scanlines correspond to observations). The orbit provides thousands of scanlines, so the few left over after partitioning are simply discarded.
3. *Weather scale.* Each patch should contain, in the cross-track direction, roughly one weather feature at a chosen scale.

Although a patch that is square on the ground would be the natural choice, it cannot be reconciled with the conditions stated above. Because the along-track sampling is about seven times coarser than the cross-track one, a patch that is square in pixels is strongly elongated on the ground, whereas a patch that is square on the ground would span far too few scanlines to characterise the striping. The geometry is therefore fixed by choosing the cross-track width to match the target weather feature scale and the along-track extent to provide enough scanlines, letting the resulting ground aspect follow.

Within these constraints, three patch sizes are defined, spanning the range of meteorological scales of interest (Table 6.1). The *mesoscale* size is adopted as the default for the denoising evaluation that follows, as it offers the best balance between spatial coverage and the number of patches available; the *local* and *synoptic* sizes are retained for, respectively, fine striping characterisation and the coverage of large weather systems.

Although the notation in Chapter 5 reports scanlines along the y -axis – the natural orientation when assembling the two-dimensional swath from the one-dimensional noise sequence – all images in this chapter are shown transposed, cross-track $\times$ along-track, so that the visualization better reflects the pixel anisotropy.

Table 6.1: The three patch sizes extracted from each orbit, in the cross-track $\times$ along-track orientation (act $\times$ alt) . Physical dimensions use the 1.22 km cross-track and 8.88 km along-track pixel spacings. The ground aspect is a consequence of the pixel anisotropy.

	Pixels (act $\times$ alt)	Physical [km]	Ground aspect	Patches/orbit
Local	116×50	142×444	1 : 3.1	1064
Mesoscale	174×100	212×888	1 : 4.2	352
Synoptic	348×200	425×1776	1 : 4.2	88

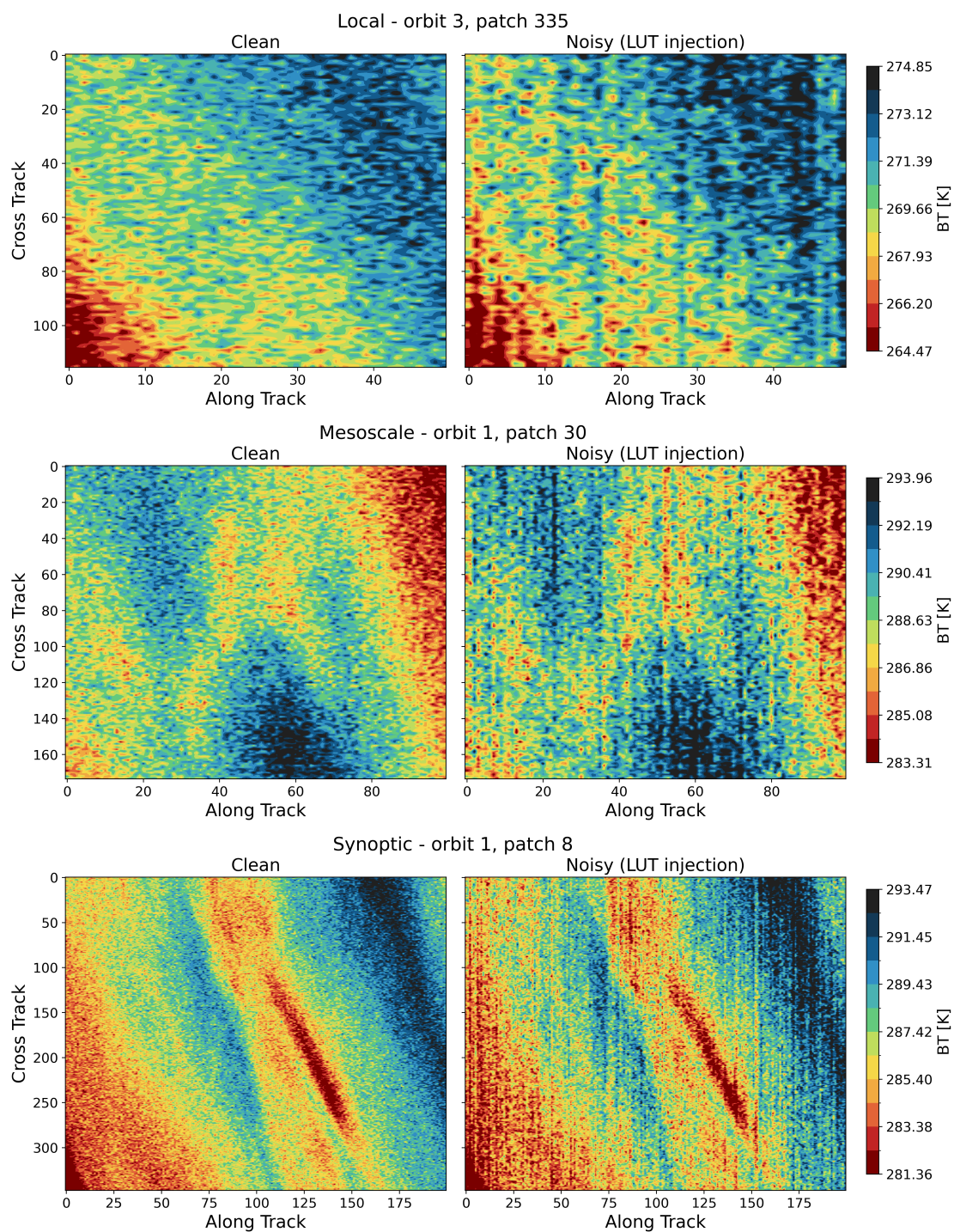


Figure 6.1: Examples of the three patch sizes: clean (left) and noise-injected (right). Striping is visible as vertical banding affecting the geophysical signal.

6.2 Methods Evaluation

This section defines the evaluation metrics and then assesses each method on the mesoscale patches, both quantitatively and through visual inspection.

6.2.1 Metrics

The destriping methods are evaluated through a combination of metrics, each capturing a different aspect of the result.

Striping Index. Of the three striping metrics introduced in Section 3.2.1, the Striping Percentage $P_{1/f}$ and Striping Ratio $R_{1/f}$ are defined on a continuous one-dimensional noise series, so they do not apply to a denoised two-dimensional image. The Striping Index (SI) is therefore the only metric retained for the evaluation. As originally introduced by Qin et al. [14], it was used for evaluation on deep space (pitch-over manoeuvre) data: a homogeneous, structureless target, for which all along-track variance is attributable to striping. On a real Earth scene this assumption fails, as the along-track variance also contains genuine geophysical structure. Exploiting the knowledge of the original scene X in the test setting, the SI is computed on the residual $R = \hat{X} - X$, which isolates what the method failed to suppress or wrongly introduced.

PSNR, SSIM. Two standard image denoising metrics are used: the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM), both computed over a fixed physical dynamic range (set to 100 K).

PSNR measures the overall reconstruction error and SSIM the structural agreement with the clean scene; for both, higher is better, with SSIM equal to 1 for identical images.

Image Blurring. An additional metric specific to destriping is the Image Distortion index (ID), introduced by Bouali and Ladjal [13]. It compares the mean cross-track power spectra of the clean scene X and the estimate $\hat{X}$ – obtained by taking the cross-track spectrum of each scanline and averaging over the scanlines – and quantifies how much the filtering has altered it:

$$\text{ID} = 1 - \frac{1}{|\mathcal{B}|} \sum_{f \in \mathcal{B}} \frac{|Q_0(f) - Q_1(f)|}{Q_0(f)}, \quad (6.2)$$

where Q_0 and Q_1 are the mean cross-track power spectra of X and $\hat{X}$, and $\mathcal{B}$ is the entire spectrum of frequencies. An index $\text{ID} \approx 1$ indicates preserved structure, while lower values flag distortion.

As will emerge in the following analysis, a characteristic failure of some methods is oversmoothing, which degrades the final product by fading the fine cross-track details of the scene. This effect is only partially captured by ID and is penalised by none of the other metrics. Therefore, a one-sided variant of the index is introduced: since blurring removes high-frequency content and thus lowers the cross-track spectral energy ($Q_1 < Q_0$), the absolute value in Equation 6.2 is replaced by $\max(Q_0 - Q_1, 0)$, penalising only a loss of energy. As the Striping Index already captures energy gains, the two metrics become complementary. To distinguish it from the original, this one-sided version is referred to as the Image Blurring index (IB) throughout the remainder of this work.

The evaluation should consider all metrics, as a single result can be misleading: Bouali and Ladjal [13] show that an aggressive low-pass filter scores well on noise-reduction measures while badly distorting the scene (caught only by the ID), and remark that the full set of indices must be read together. For this reason, the quantitative metrics are complemented throughout by visual inspection of the denoised image and its residual: Dabov et al. [19] indeed assess BM3D by both PSNR and subjective visual quality.

6.2.2 Per-Method Analysis

Each method is assessed both visually and through the metrics. The visual assessment relies on a four-panel layout: the clean scene and its striped counterpart on the top row, the destriped estimate alongside the residual $\hat{X} - X$ on the bottom. An ideal method would produce a null residual, so any vertical banding indicates striping left unremoved, while any geophysical structure is signal mistakenly erased.

Wiener Filter

The knowledge of a precise estimate of the noise PSD does not bring the expected advantage. As shown in the denoised patch of Figure 6.2, the filter affects the geophysical structure (scanlines 0–10 and 40–60) while leaving the striping pattern throughout.

The Wiener filter is designed to suppress the frequencies at which the noise spectrum dominates the signal one. As established in Chapter 3, the flicker component carries increasingly more energy as the frequency approaches zero: where scene variation shares that same spectral band, the filter cannot tell them apart and attenuates the signal together with the noise. Indeed, the residual shows no sharp scene features but rather a slow, low-frequency variation – the scene structure the filter has removed.

The window length is set to $N_{\text{win}} = 5\,000$ samples ($\Delta f \approx 0.5$ Hz), but the result is

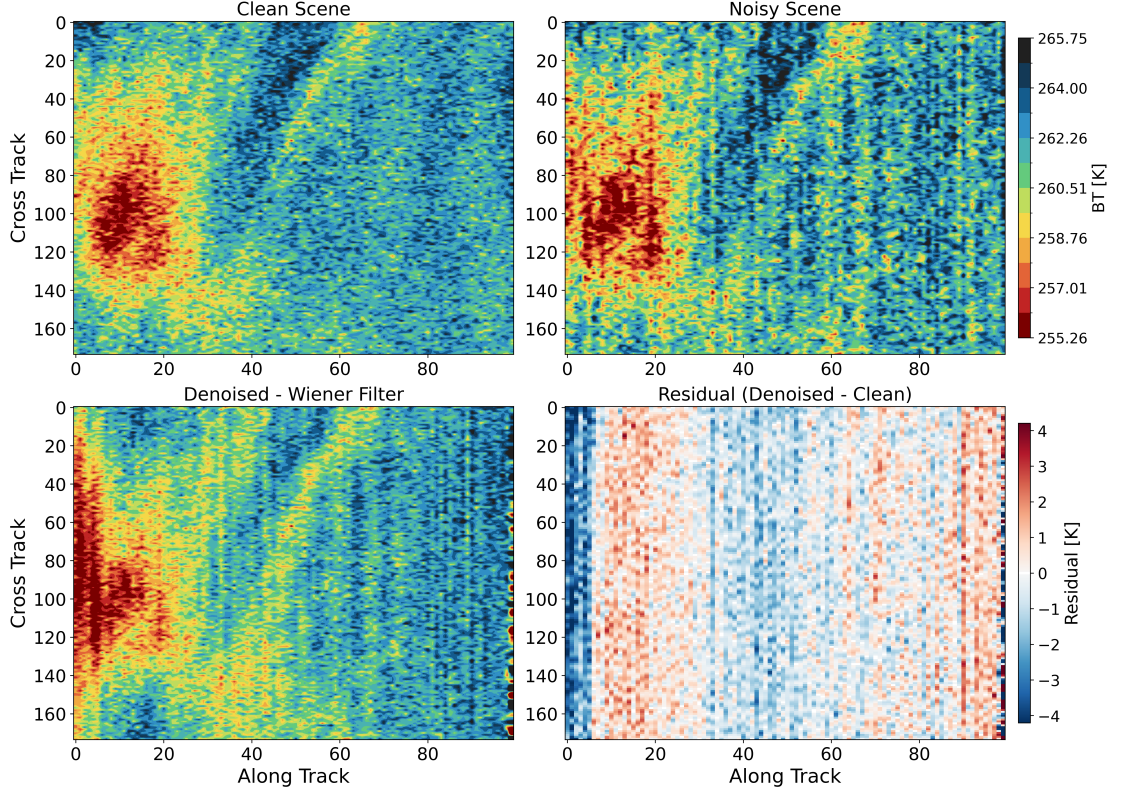


Figure 6.2: Wiener Filter on a mesoscale patch (orbit 1, patch 315). The filter distorts the scene structure yet leaves the striping pattern, failing on both fronts.

essentially insensitive to it: shorter or longer windows do not improve the outcome, since the limitation is the spectral overlap of noise and signal rather than the frequency resolution. The metrics confirm the visual assessment: the PSNR falls ($38.60 \rightarrow 37.05$) and the Striping Index rises ($1.47 \rightarrow 2.02$), as the residual has more along-track than cross-track variance; the IB decreases ($0.99 \rightarrow 0.96$), exposing the loss of energy in the cross-track direction, the very direction in which the filter operates. Two metrics, by contrast, report a misleading improvement: SSIM and ID rise ($0.91 \rightarrow 0.94$ and $0.82 \rightarrow 0.95$ respectively), as the removal of the noise energy offsets the loss of the signal attenuated with it.

The filter thus fails on both fronts it is meant to address, leaving a distorted scene with striping patterns still visible; therefore, it is not retained as a viable destriping method.

The negative outcome is itself a finding: precise PSD knowledge does not suffice when stripe and scene share a spectral band, implying that spatial structure – not spectral characterisation alone – should also be exploited.

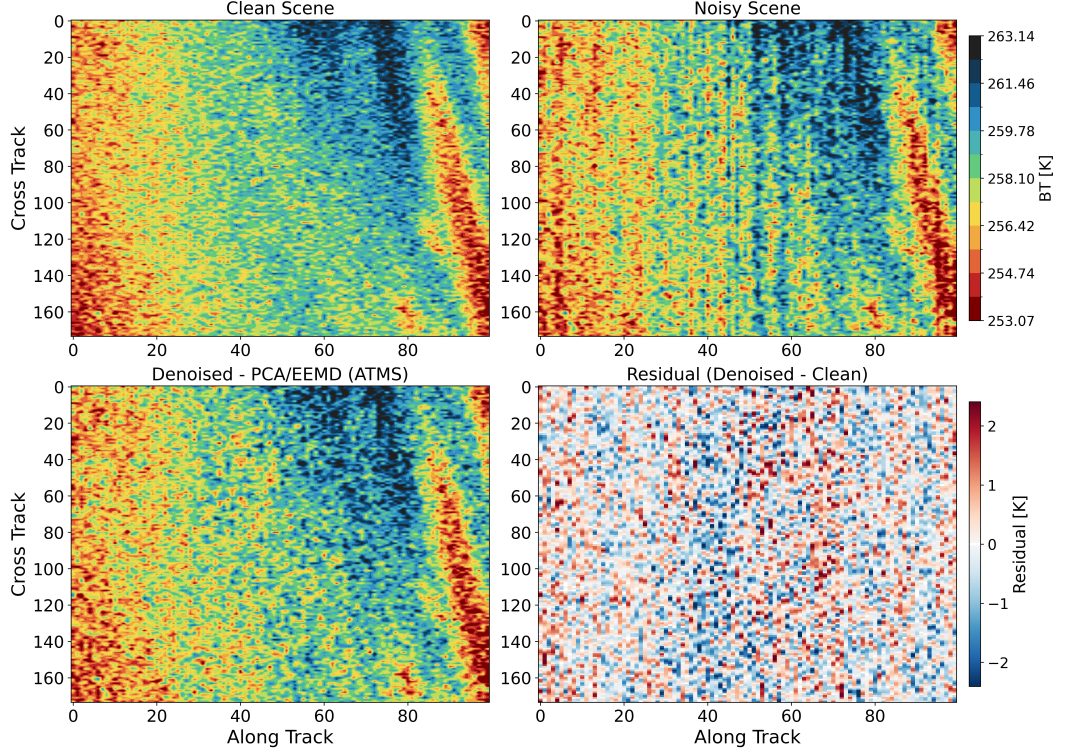


Figure 6.3: PCA/EEMD (ATMS approach) on a favourable scene (orbit 3, patch 351). The striping is removed and the residual is dominated by incoherent noise, with the geophysical structure relatively well preserved.

PCA/EEMD

Throughout this evaluation and the rest of the work, the two PCA/EEMD configurations introduced in Section 5.2 are referred to as the *ATMS* and *GMI approaches*. These names are only a shorthand to designate the methods explored in the two different studies from Qin et al. [14] and Dong and Zou [16] on ATMS and GMI respectively; neither is specific or uniquely applied to the corresponding instrument.

As the operational benchmark for microwave radiometers, the combination of PCA and EEMD performs well on favourable scenes, where the two approaches behave comparably.

Figure 6.3 shows such a case for the ATMS approach; the GMI result is near-identical and is omitted. Both reduce the Striping Index towards unity ($1.56 \rightarrow 1.06$ for ATMS, $1.56 \rightarrow 1.17$ for GMI) and raise the PSNR (+1.65 and +1.29 dB), leaving a residual mostly dominated by incoherent salt-and-pepper noise. The IB stays at 0.99 throughout, confirming this method is not subject to oversmoothing.

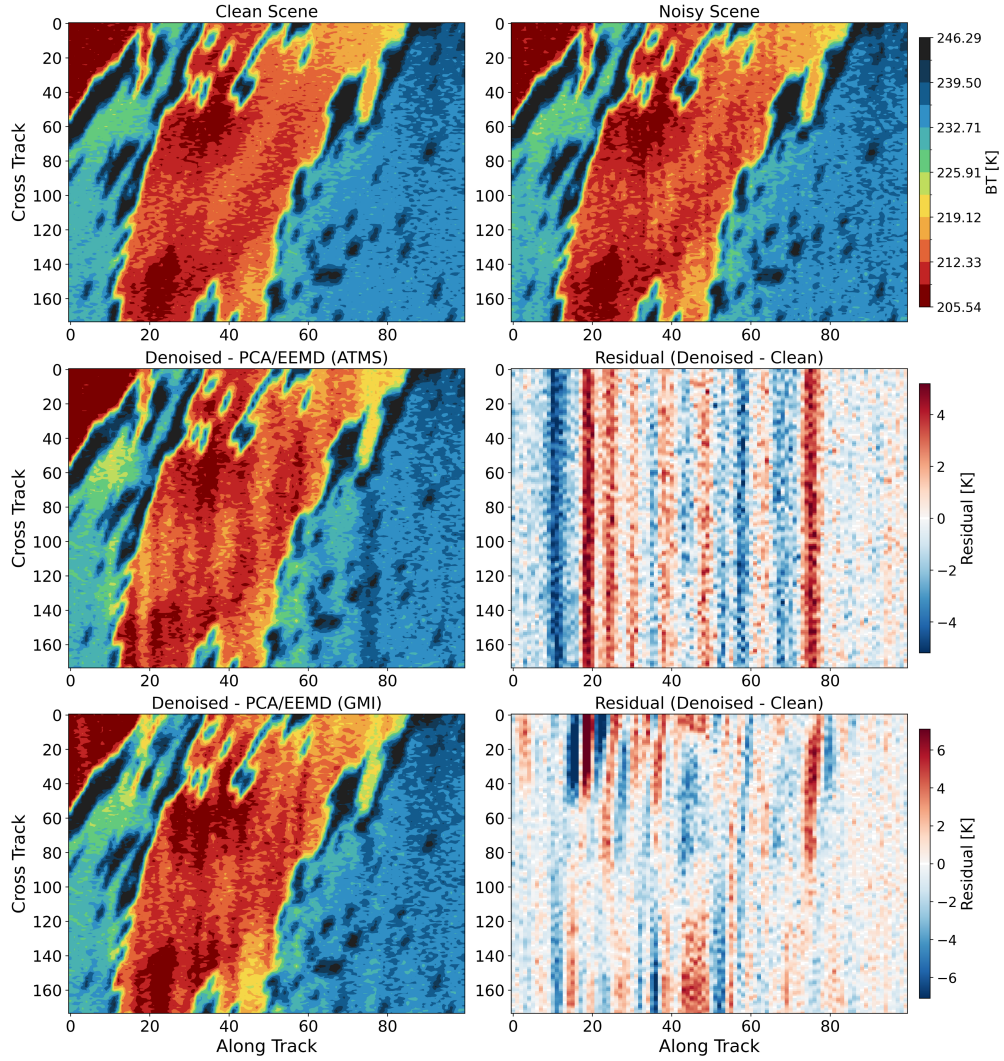


Figure 6.4: Failure of PCA/EEMD on a structured scene (orbit 3, patch 323). *(Top)* Clean and Striped patch. *(Middle)* ATMS, which strips out band-like structure. *(Bottom)* GMI, which removes localized features.

The performance is, however, strongly scene-dependent.

On scenes where the leading principal components conflate the stripe offset with genuine structure, features are removed along with the IMFs. Neither approach is consistently better, but the two fail in characteristically different ways, as illustrated in Figure 6.4. The ATMS approach (*Middle*) removes modes from a single dominant component, so when it over-corrects it strips out band-like structure (scanlines 20–80): this is penalised by both the Striping Index and the PSNR (SI 1.39→5.14, PSNR −5.67 dB). The GMI configuration, instead, removes a single mode from

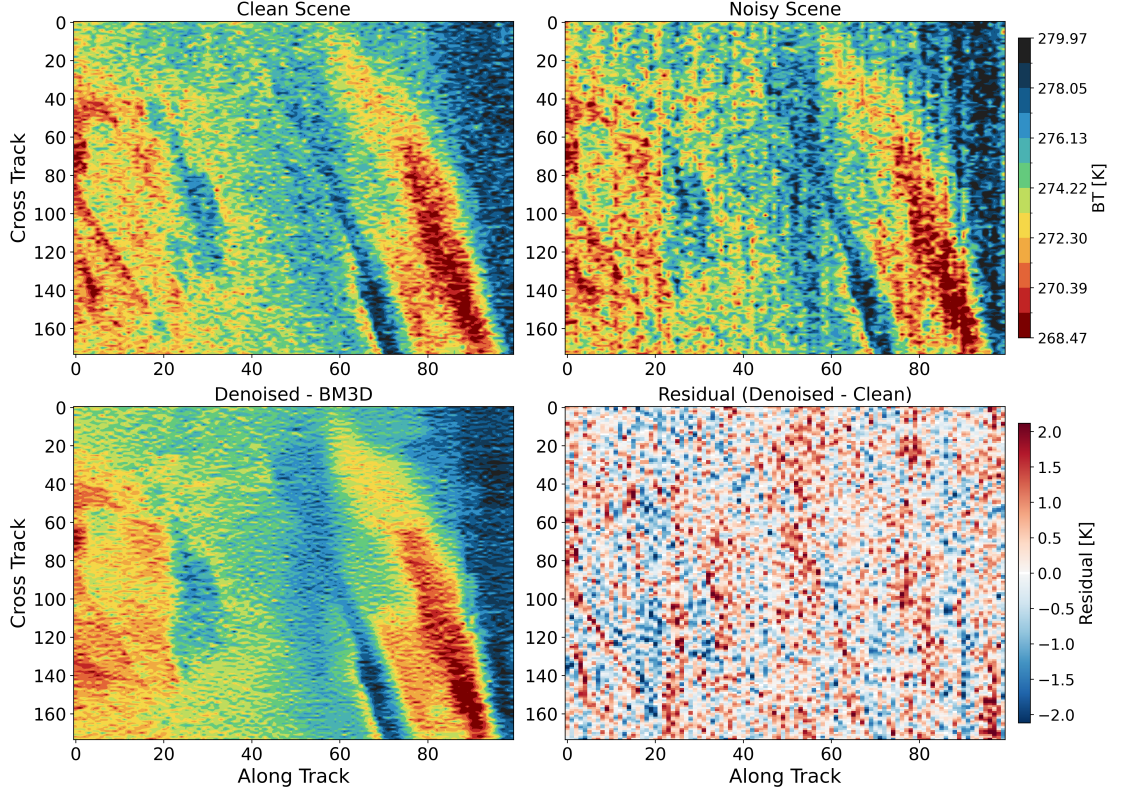


Figure 6.5: BM3D on a mesoscale patch (orbit 3, patch 23). Both fine details and partial stripes are visible in the residual.

each of the first three coefficients, so the residual structure is more localized and less directional (scanlines 10–50): the PSNR registers the damage (-6.27 dB) while the Striping Index actually improves ($1.39 \rightarrow 1.22$).

This contrast exposes a limitation of the Striping Index: it only detects coherent, directional distortion, so it captures the ATMS failure but misses the GMI one, which only the PSNR reveals. In both cases, the IB shows only slight decrease.

BM3D

The BM3D algorithm oversmooths the scene in the cross-track direction. As shown in Figure 6.5, fine features (scanlines 0–40) are blended together and the striping is not specifically targeted. The block geometry (H, W) – the along-track and cross-track extent of each block – is chosen to span an approximately square area on the ground. The cross-track extent is taken as the full patch width ($W = 174$), so that BM3D has the widest possible set of fragments to match against. The along-track extent follows from the pixel geometry: a square ground footprint

would require $174 \times 1.22/8.88 \approx 23.9$ pixels, rounded to $H = 23$ to align with BM3D’s internal 8×8 reference block and stride of 3, matching the full block length exactly.

BM3D’s failure is the clearest illustration of why no single metric suffices: both SI and PSNR show improvement ($1.47 \rightarrow 1.11$ and $+3.48$ dB), as the residual is non-directional and oversmoothing minimises the mean-square error, so on these two it would appear excellent. The IB, however, collapses ($0.99 \rightarrow 0.82$), exposing the blurring that neither SI nor PSNR register. The ID barely decreases ($0.84 \rightarrow 0.82$), motivating the implemented one-sided variant.

Reducing the strength of the filtering by tuning the internal parameters of the algorithm – the hard-threshold multiplier of the first stage and the Wiener regularization of the second – leads to a milder smoothing yet leaves more striping, so no parameter setting removes the banding while preserving details.

Hybrid Method

The behaviour of the Hybrid is compared against the three individual methods on two representative patches.

On a favourable scene (Figure 6.6, Table 6.2) the Hybrid gives the best overall result. BM3D still attains the highest PSNR (42.05 dB) but its IB collapses to 0.85, flagging the vertical streaking visible in the patch; the ATMS configuration adds cross-track structure (scanlines 60–80) and the GMI approach visibly removes features (the red area at scanline 60). The Hybrid avoids all three failures: its PSNR (41.90 dB) matches BM3D, its SI (1.16) is close to the best (1.09), and its IB stays at 0.95. As already stated in the PCA/EEMD evaluation, the SI misleadingly rewards GMI as its structure removal is non-directional; nonetheless, the Hybrid does remove striping and the SI correctly evaluates it.

On an unfavourable scene (Figure 6.7, Table 6.3) BM3D is instead the best, both visually and according to the SI and PSNR, while the Hybrid’s limitation becomes apparent. Its first stage is PCA/EEMD, so any spurious cross-track structure introduced by the removal of the single IMF is already present in the image BM3D receives. The Hybrid can only refine the output of its first stage, not correct it: this is its intrinsic limitation.

Given the distinct role of the two combined methods, two alternatives to the proposed configuration are inspected.

The first is whether the BM3D stage is needed at all: since using a single IMF avoids the over-removal, one might apply PCA/EEMD in this setting alone. This,

as expected, proves to be insufficient: on the favourable scene it gives a PSNR of 39.50 dB, whereas adding BM3D raises it to 41.90 dB, while the SI is essentially unchanged (1.13 against 1.16).

The second is the order of application: reversing it so that BM3D precedes PCA/EEMD. This brings no advantage: on favourable scenes the performance is slightly worse (SI 1.25 against 1.16, PSNR 41.30 against 41.90 dB), while on unfavourable ones the two are practically indistinguishable (SI 1.71 against 1.69, PSNR 40.44 against 40.52 dB).

Across both scenes the IB separates the methods: in particular, it isolates BM3D’s oversmoothing, which the PSNR and SSIM reward. On the other hand, the two-sided ID barely distinguishes the methods and over-rewards the Hybrid even in the case of a bad performance (bottom lines of Tables 6.2, 6.3) . This confirms the value of the one-sided variant.

Overall, as quantified in Section 6.3, the Hybrid is the most robust of the methods: where the individual methods succeed it matches or visually surpasses the best of them, and where they fail it performs better; its only limitation is the structure lost or altered by its PCA stage, which no subsequent step can recover.

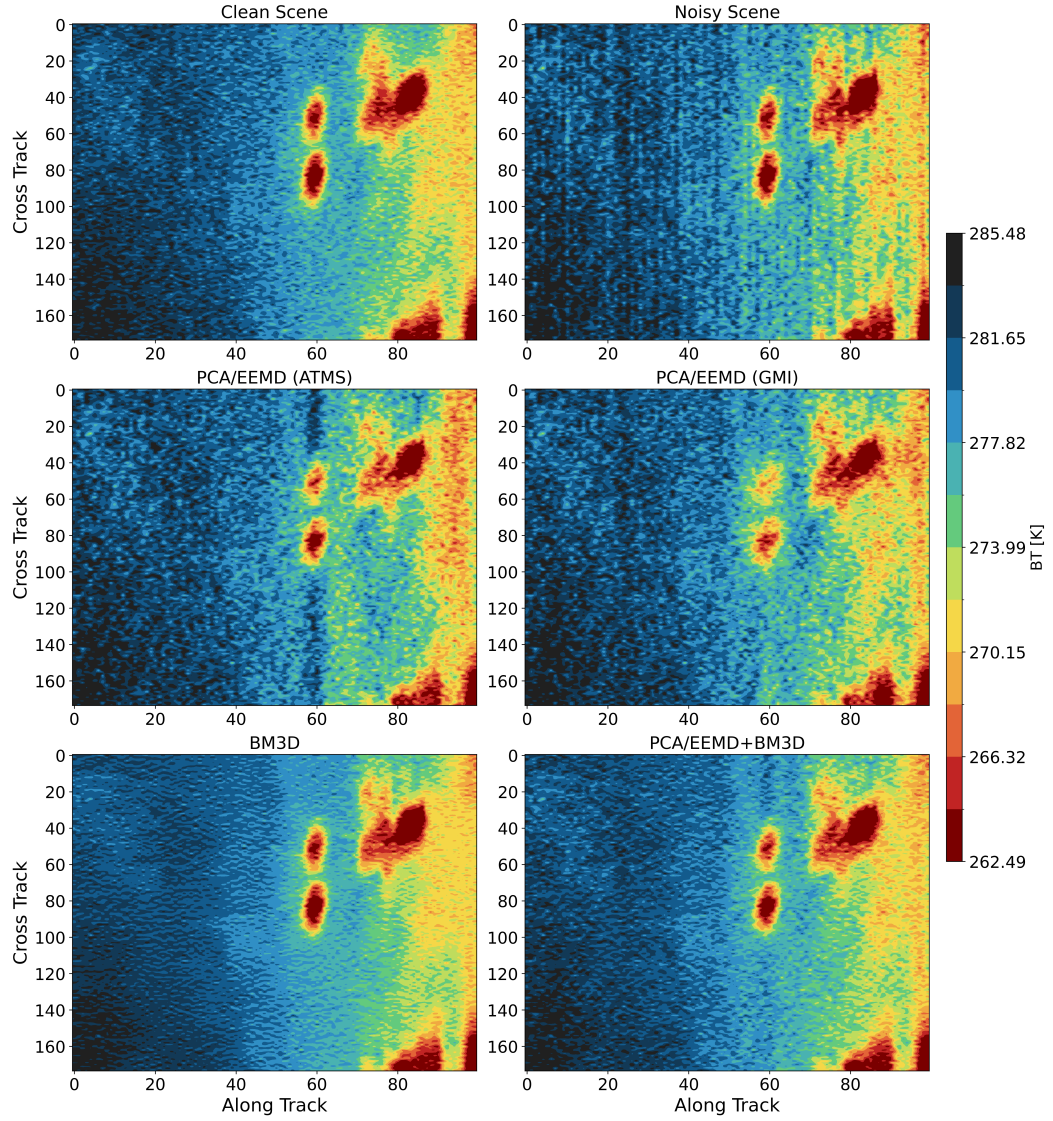


Figure 6.6: Method comparison on a favourable patch (orbit 3, patch 52).

Table 6.2: Metrics of the favourable patch of Figure 6.6.

	Noisy	ATMS	GMI	BM3D	Hybrid
SI	1.47	2.09	1.09	1.11	1.16
PSNR [dB]	38.38	36.88	38.13	42.05	41.90
SSIM	0.91	0.91	0.92	0.96	0.96
IB	0.99	0.99	0.99	0.85	0.95
ID	0.86	0.86	0.86	0.85	0.93

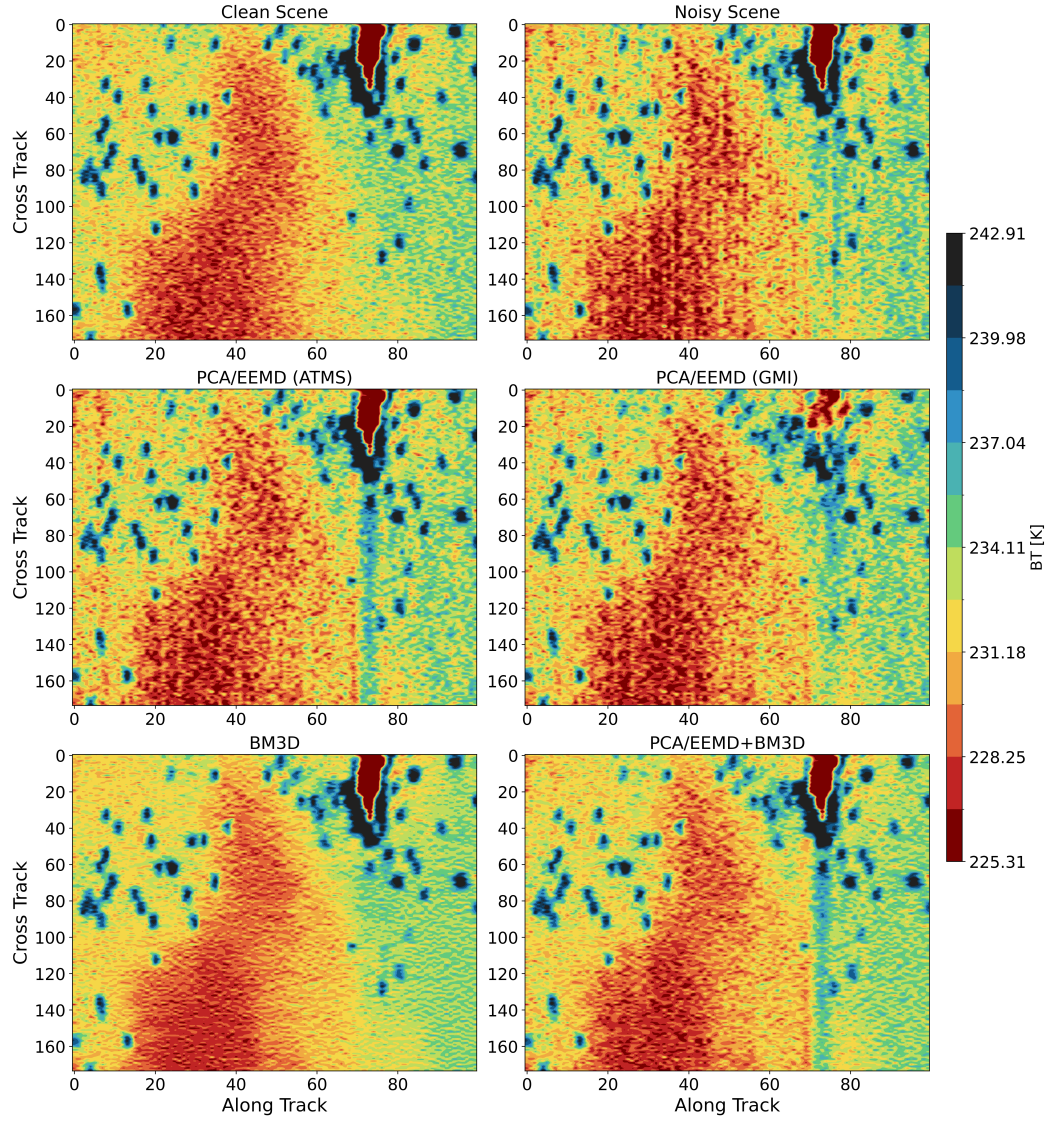


Figure 6.7: Method comparison on an unfavourable patch (orbit 3, patch 325).

Table 6.3: Metrics of the unfavourable patch of Figure 6.7.

	Noisy	ATMS	GMI	BM3D	Hybrid
SI	1.37	1.52	1.20	1.12	1.69
PSNR [dB]	39.64	39.19	35.76	42.02	40.52
SSIM	0.94	0.94	0.92	0.96	0.96
IB	0.99	0.99	0.98	0.89	0.95
ID	0.91	0.91	0.91	0.89	0.94

6.3 Challenge Baseline

The preceding analysis showed that no single metric is sufficient, as each captures one aspect of the result but is blind to at least one failure mode:

- *SI*. It correctly assesses residual striping through the known clean scene X , but is misled when the removed structure is non-directional.
- *PSNR*. It strongly penalises a method that distorts the signal but rewards oversmoothing, so BM3D always attains the highest PSNR.
- *SSIM*. It behaves like the PSNR but, already close to 1 for the noisy image, adds little discriminating power.
- *ID*. It does not separate genuine denoising from oversmoothing, and rewards visually degraded results; like the SSIM, it barely moves across methods.
- *IB*. It addresses the oversmoothing shortcoming specifically: it stays essentially unchanged for methods that do not blur the fine cross-track details of the scene and decreases as the effect worsens. It is a targeted metric, of little value on its own but useful as a complement to the others in assessing a complete denoiser.

For these reasons, neither the SSIM nor the ID is retained for the evaluation, while the three remaining metrics are combined into a single score, as described next.

6.3.1 Scene-Adaptive Score

As asserted in Chapter 1, the ultimate objective of this work, and of the challenge in particular, is to remove striping while preserving the genuine geophysical signal. Although none of the metrics works individually, a single automatic score is still required for the challenge, since the visual inspection relied on throughout this work cannot be carried over to that setting. To meet this need, the metrics already in use are combined into a single index, designed to reflect the visual assessment and to measure precisely this dual goal. For this reason it is referred to as the *Striping-Preservation Index* (SPI).

The metrics are combined by weighting each one according to the scene, since their reliability is itself scene-dependent:

- On a near-uniform scene, striping is the dominant artefact and the SI assesses it most reliably, so it should carry the most weight.

- On a strongly varying scene, preserving the genuine structure is critical, yet it is precisely in these cases that the SI is least reliable: non-directional structure removal goes undetected by it. The PSNR fills this gap, as it directly penalises any distortion of the signal regardless of its directionality.
- In all cases, oversmoothing must be penalised, as it neither removes striping nor preserves the scene: this is the role of the IB.

The SPI is therefore computed in three steps:

1. *Scene classification.* The scene complexity is quantified by a parameter $t \in [0,1]$: first, the clean scene X is high-pass filtered, then its standard deviation is computed and normalised between the 5% and 95% percentiles over the dataset ([1.31, 4.81] K). A low value corresponds to a smooth, near-uniform scene, while a high value to a richly textured one.
2. *Metrics normalization.* Each metric is first mapped to a goodness value $g \in [0,1]$ so that the three can be combined despite their different natural scales. The SI, whose ideal value is 1 and which may deviate in either direction, is mapped by a symmetric exponential $g_{\text{SI}} = \exp(-|\text{SI} - 1|/s)$, where $s = 1.2$, peaking at $\text{SI} = 1$.
The PSNR, being unbounded and monotonic, and the IB, bounded but compressed near its ideal value of 1, are both mapped by clipped linear functions between empirical lower and upper bounds, [25, 45] dB and [0.75, 1] for g_{PSNR} and g_{IB} respectively.
3. *Score computation.* The score is finally computed through the weighted sum $\text{SPI} = w_{\text{SI}} \cdot g_{\text{SI}} + w_{\text{PSNR}} \cdot g_{\text{PSNR}} + w_{\text{IB}} \cdot g_{\text{IB}}$ with weights

$$\begin{aligned} w_{\text{SI}} &= 0.50 - 0.45 t \\ w_{\text{PSNR}} &= 0.35 + 0.55 t \\ w_{\text{IB}} &= 0.15 - 0.10 t \end{aligned}$$

where t is the scene complexity. The weights interpolate linearly between the two regimes: on a smooth scene ($t = 0$) the SI dominates, on a textured one ($t = 1$) the PSNR takes over, while the IB acts throughout as a small guard against oversmoothing. The interpolation endpoints were tuned on representative patches.

The reliability of the SPI was tested on several patches; here, it is only checked against the two examples analysed in Section 6.2.2, where the visual assessment and the individual metrics were already discussed.

On the favourable case (Figure 6.6) the Hybrid attains the highest SPI (0.850), matching the visual evaluation. The scene carries only some structure ($t = 0.17$): if the SI completely dominated, the GMI approach would have mistakenly placed first; the contribution of the PSNR flags it, correcting the GMI score down to 0.815. The IB, in turn, prevents this correction from over-rewarding BM3D, bringing it to 0.816, just above GMI but well below the Hybrid. The ATMS configuration falls last (0.568), having the worst SI and PSNR.

The example shows why, even on a low-structure scene, the PSNR and IB must retain part of the weight rather than deferring entirely to the SI.

On the unfavourable patch (Figure 6.7) the ranking again matches the visual estimation: BM3D scores highest (0.840), followed by the ATMS approach (0.722) and the Hybrid (0.703) – both penalized for the structure altered by PCA/EEMD – and the GMI approach last (0.697), because of its lower PSNR. As the scene has higher complexity ($t = 0.33$), the weighting shifts towards the PSNR, but the SI still retains a non-negligible weight: although on a structured scene its reading is biased, the penalization of directional structure loss is still useful for correct ranking. The IB weight is slightly smaller, as its discriminative power decreases with scene structure.

As a failure reference, the Wiener filter – discarded in Section 6.2.2 – scores 0.327 and 0.396 on the two patches respectively, confirming that the score resolves the full range from failure to success rather than saturating near the top.

6.3.2 Baseline Evaluation

The four methods are evaluated with the SPI over the three available orbits (1056 mesoscale patches), with no train–test division: being non-learned, the methods are scored on the whole dataset. Figure 6.8 reports the per-method distributions, which are read on three levels: the median for typical performance, the interquartile range for consistency, and the outliers for failure modes.

On the first two aspects the Hybrid PCA/EEMD+BM3D stands out. It attains the highest median SPI (0.857) and a tight interquartile range, so it is both the best on a typical scene and among the most consistent. The other methods, instead, show a weakness in at least one of the two: the ATMS configuration reaches a fair median (0.824) but with by far the widest spread, failing badly and often; BM3D is tightly concentrated but sits lowest in the median (0.814), as oversmoothing penalises it on every scene; the GMI configuration lies in between, with a good median (0.841) and a moderate spread but a heavy low-score tail.

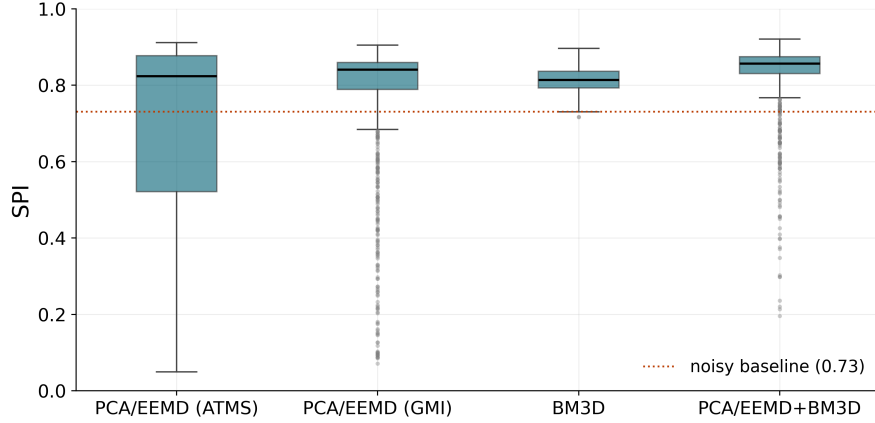


Figure 6.8: SPI distribution of the four methods over the three orbits (1056 mesoscale patches). The Hybrid has the highest median and a tight spread; its low-score outliers collect the high-structure scenes where the PCA stage removes genuine signal.

The Hybrid is therefore proposed as the challenge baseline against which participant submissions should be measured. The choice of a non-learned method as baseline not only provides a fair reference but also makes it possible to establish whether machine-learning destriping methods bring a genuine advantage.

An analysis of the rankings may at first seem to contradict this result: the Hybrid is among the top two methods on 88% of patches, yet it comes first on only 39%, being mostly outscored by the PCA-based configurations. This is a consequence of the scene distribution, which is dominated by low-structure patches (median structure level 1.52 K, with the most structured patch reaching 7.9 K). On such scenes the methods differ only slightly, so the ranking is decided by negligible amounts: the Hybrid loses by a median of just 0.018 to the ATMS configuration and 0.008 to GMI, on scenes of median structure 1.33 K and 1.40 K respectively. Where the Hybrid wins, by contrast, the median margins are 0.186 and 0.019, about ten times larger. The Hybrid thus performs comparably on the easy scenes and gains decisively on the hard ones, which is exactly the robustness its tight interquartile range reflects.

The per-patch running time, measured during the same evaluation, confirms the cost asymmetry between the methods: the PCA/EEMD configurations are markedly cheaper than BM3D and the Hybrid, at 0.17 and 0.52 s/patch against 4.6 and 4.7 s/patch, hence minutes against hours of total running time over the dataset.

The Hybrid’s one genuine weakness is the tail of low-scoring outliers visible in Figure 6.8: 122 patches (11.6%) fall below the lower whisker. These are not scattered failures but a coherent group of high-structure scenes, with a median structure level of 4.10 K. On every one of them BM3D outcores the Hybrid (0.831 against 0.645 median SPI); the other methods perform even worse (0.453 and 0.288 median SPI for GMI and ATMS), confirming that the failure originates in the PCA stage. The mechanism is the one identified in Section 6.2.2: where the scene is richly textured, the principal components conflate genuine cross-track structure with the stripe mode and remove it, and BM3D cannot restore what is already lost (Figure 6.9). The GMI approach, based on the same decomposition, suffers the failure more severely: its tail is both larger (163 patches, 15.4%) and deeper (median SPI 0.466 against the Hybrid’s 0.645), so the Hybrid contains the high-structure failure better than the benchmark it builds on.

Although these scenes are a minority, they are nonetheless important for NWP, motivating the refinement proposed in the next section.

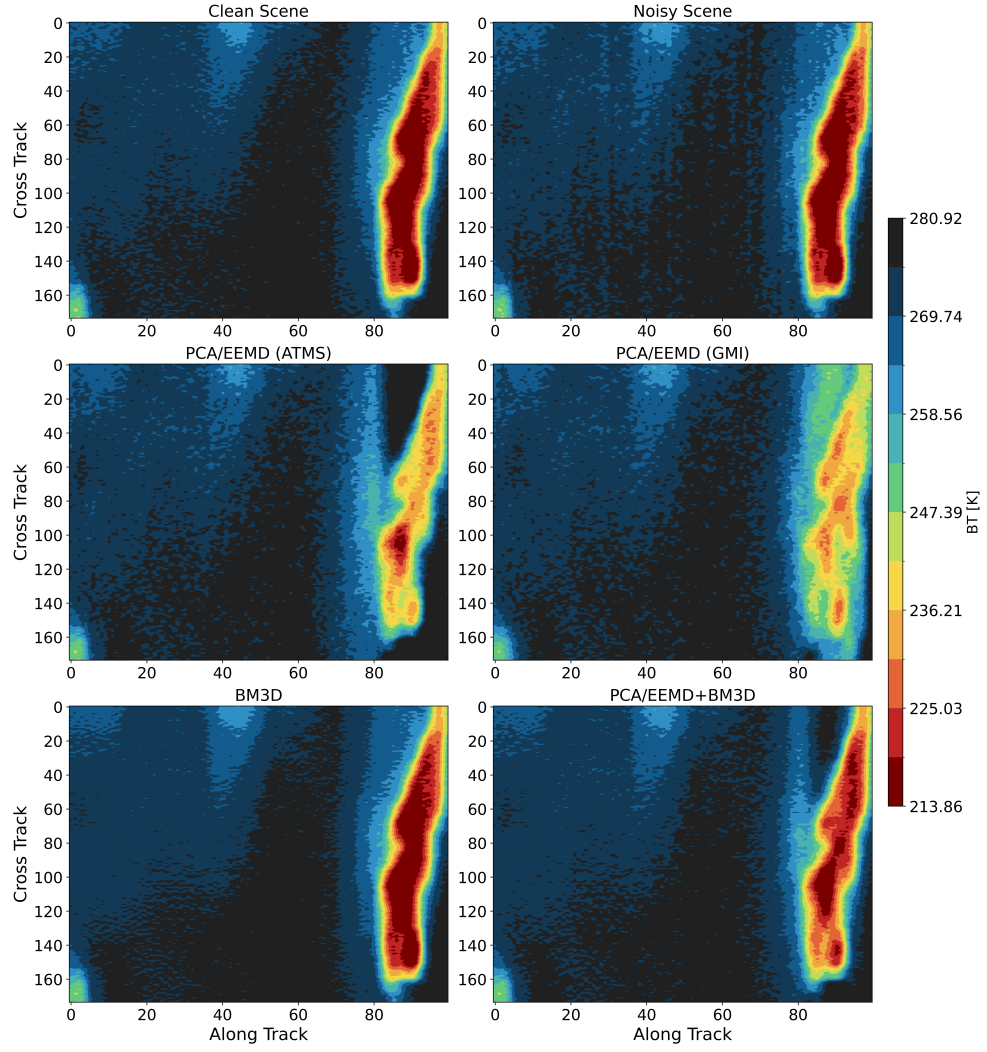


Figure 6.9: A worst-case patch from the Hybrid’s tail (orbit 3, patch 60; structure level 6.08 K, well above the dataset median). The PCA stage removes genuine cross-track structure, which the BM3D stage cannot recover.

Table 6.4: Metrics and SPI of the tail patch of Figure 6.9.

	Noisy	ATMS	GMI	BM3D	Hybrid
SI	1.41	63.61	1.20	1.15	26.64
PSNR [dB]	39.29	22.55	23.58	42.36	28.72
IB	0.99	0.99	0.98	0.91	0.96
SPI	-	0.05	0.08	0.86	0.21

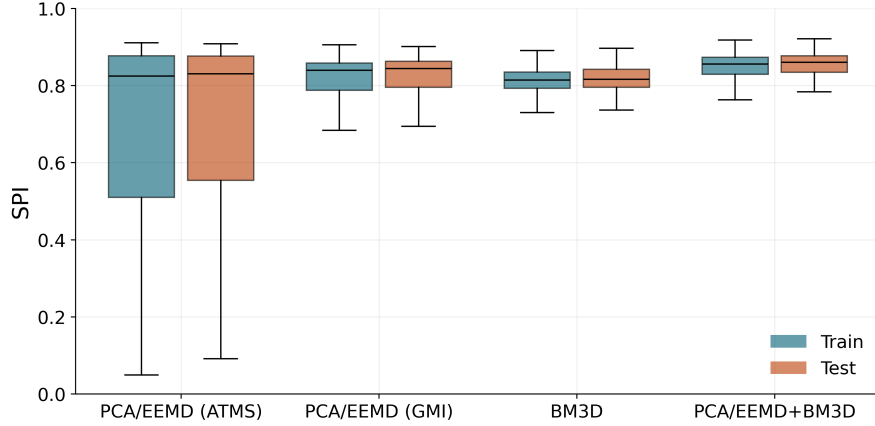


Figure 6.10: SPI distribution on the training and test sets for each method, under the 80/20 stratified split. The matched distributions confirm the split is balanced.

6.4 Train–Test Split

The challenge dataset must be divided into training and test sets that share the same distribution of scenes, so that performance on one transfers to the other. The split is performed exploiting the scene classification from the SPI: the patches are first stratified by the complexity value t into three equally populated bins and each bin is split with an 80/20 ratio; the training and test sets are then obtained by recombining the respective portions, yielding 846 and 210 patches.

Because the baseline methods are not learned, they provide a direct check of fairness: equal scores on the training and test sets confirm that the complexity distribution is preserved across the two. Figure 6.10 reports, for each method, the SPI distribution on the training and test sets side by side. The two are closely matched for every method ($\Delta = 0.002\text{--}0.006$), confirming that this approach can be used for the challenge.

6.5 A Structure-Aware Hybrid

As diagnosed in Section 6.3.2, the baseline’s residual failure is confined to high-structure patches: in such cases the PCA stage removes genuine signal, whereas BM3D would have been the safer choice. If these patches could be identified in advance, they could be routed to BM3D alone. The difficulty is that the failure is defined through the clean scene X , while only the acquired noisy swath Y is available operationally.

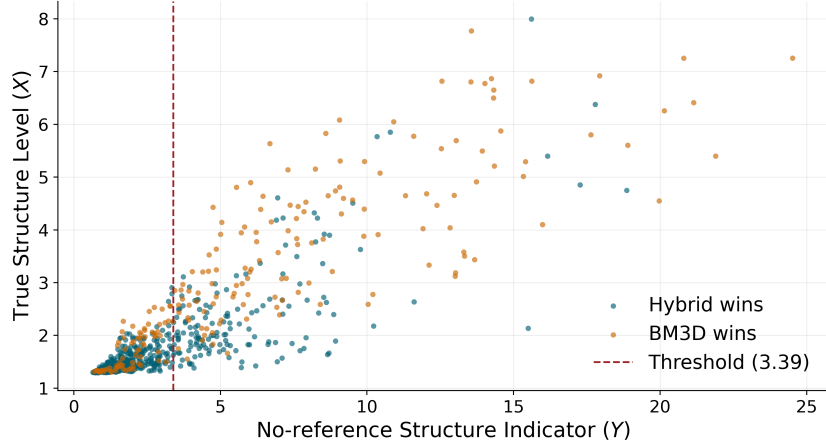


Figure 6.11: No-reference structure indicator against the reference complexity t on the training set, coloured by the better-performing method. The two are strongly correlated ($\rho = 0.91$), and the calibrated decision threshold is marked.

A no-reference indicator of scene structure is therefore defined on Y , which must handle both the thermal and the striping noise, to prevent either from inflating the complexity value of the scene.

The indicator is computed in three steps. First, each scanline is smoothed by a low-pass Gaussian filter, suppressing the incoherent thermal noise. Then the scanline mean is subtracted, treating the striping as a nearly constant per-scanline offset; this holds as an approximation, especially for the MWI striping, which is non-periodic. Finally, the standard deviation of the result is taken as the indicator: once the thermal noise and the stripe offset have been attenuated, the remaining variation can be attributed to genuine scene structure.

The indicator is validated against the true scene complexity on the training set. Figure 6.11 shows that it correlates strongly with the reference complexity t (Spearman $\rho = 0.91$), confirming that it captures the same information without access to the clean scene X .

The refinement of the Hybrid method is then straightforward: a patch whose no-reference structure indicator exceeds a threshold is processed by BM3D alone, otherwise the standard Hybrid applies. The threshold is calibrated on the training set as the value that maximises the mean routed score; it is set to $r = 3.39$, which routes 33% of the training patches to BM3D.

Evaluated on the test set (Table 6.5), the Structure-Aware Hybrid (SA-Hybrid) improves on the plain Hybrid. The median is essentially unchanged (0.860 against 0.861), while the mean rises markedly (0.854 against 0.833), confirming the targeted

tail improvement. The running time is comparable (1.7% faster), since the BM3D stage, not the skipped PCA one, dominates the Hybrid’s cost.

Beyond the refinement itself, which is limited to high-structure patches, the result carries a more general implication: scene complexity can be estimated from the noisy data alone, allowing the adaptation of the destriping strategy per scene.

Table 6.5: Final SPI on the test set (210 patches): BM3D, PCA/EEMD+BM3D and the Structure-Aware Hybrid.

	Median SPI	Mean SPI
BM3D	0.816	0.818
PCA/EEMD+BM3D	0.861	0.833
SA-Hybrid	0.860	0.854

Chapter 7

Conclusions

This thesis was motivated by the problem of striping noise affecting microwave satellite instruments and its consequences for downstream applications. As the artefact is intrinsic to these sensors and may worsen over a mission’s lifetime, it must be addressed at the image level. To accelerate progress, the work was framed as the preparation of a scientific challenge; its two objectives – building a realistic dataset and establishing a baseline against which submissions can be measured – are revisited here, before the remaining limitations and next steps are discussed.

7.1 Summary of Contributions

The first objective was met by characterising the MWI instrument noise from the ground-test data. The noise was modelled as a Gaussian, two-component process – a white floor and a long-memory flicker component – with a reproducible power spectral density, treated as wide-sense stationary. Its spatial manifestation as striping was then studied through the operational calibration chain, confirming that calibration mitigates but cannot remove the artefact, which must therefore be processed at the image level. On this basis, two synthetic generators were developed and validated against the observed noise: a parametric Thermal+ $1/f$ model and an empirical Look-Up Table model. The dataset was built by injecting the LUT noise – as it reproduces the spatial appearance of the striping more faithfully – into MWI Level-1B scenes.

As for the second objective, four destriping methods – the Wiener Filter, Principal Component Analysis with Ensemble Empirical Mode Decomposition (PCA/EEMD), Block-Matching and 3D Filtering (BM3D) and a Hybrid combining the last two – were implemented and evaluated on the generated dataset. No single metric could correctly rate the overall performance of the methods, which led to the definition

of the Striping-Preservation Index: a scene-adaptive score designed to reflect the dual goal of removing striping while preserving the genuine signal. Scored across the dataset, the Hybrid PCA/EEMD+BM3D emerged as the most robust method, combining the highest median score with a tight spread, and was therefore proposed as the challenge baseline. Its one residual weakness – a loss of structure on the most complex scenes, inherited from its PCA stage – was diagnosed and addressed by a structure-aware refinement, identifying such scenes from the noisy data and routing them to BM3D alone.

More than any single method, the contribution of this evaluation is to have exposed where the operational benchmark fails and to have indicated, through both the Hybrid and the adaptive routing, the direction in which a more robust denoiser can be built.

7.2 Limitations and Future Work

This work remains a preliminary study: it tests a range of destriping methods and confronts the central difficulty of removing striping without corrupting the underlying signal, but several aspects remain open and should guide its continuation:

- *Noise characterisation and dataset.* The noise model is built on ground-test data that are limited in both the length and the number of the observations. Longer continuous measurements would tighten both the spectral and the Allan deviation estimates and, in turn, let the synthetic generators – the LUT model in particular – generalise more reliably. Indeed, this was also recommended to Agencies by the MWI Science Advisory Group in view of future missions.
The challenge dataset itself, currently restricted to three simulated orbits, should likewise be enlarged and, once MetOp-SG B1 is in orbit, extended to real measurements.
- *Scene-adaptive score.* The SPI captures the intended striping–preservation duality, but its weight and goodness-mapping ranges were tuned on representative patches. A larger and more systematic validation is needed before it can be fixed as the official challenge metric.
- *Testing scope.* The evaluation was limited to mesoscale patches with LUT-injected noise. It should be extended to the other patch sizes and, in particular, to varying noise levels, so as to assess the robustness of the methods under degraded operational conditions.
- *Multispectral destriping.* The present work treats the single MWI-13 V channel in isolation. Exploiting the multispectral information across channels should

help disentangle the geophysical signal – which shows some spectral correlation – from the flicker noise, which is spectrally uncorrelated across channels that do not share common electronics.

- *Runtime.* The PCA/EEMD configurations are markedly cheaper than BM3D and the Hybrid. For a large challenge dataset, and especially for operational use, this cost matters; an optimised BM3D stage or a faster alternative is worth exploring.
- *Baseline definition.* The Hybrid is proposed as the baseline, but the final challenge specification – the dataset, the scoring metric and the baseline itself – will be defined together with ESA.

These define the path from the present preparatory work to a released challenge. The real test will then be the impact on end-user applications: assimilation experiments comparing the original and variously destriped products in NWP, for instance, would ultimately measure the effectiveness of the proposed methods. MetOp-SG B1 will provide the data and the research community the methods to exploit it. Together, they can let the MWI and the other instruments on board deliver the improvement they were designed for, supporting the progress of Earth observation.

Bibliography

- [1] Andrew J. Tatem, Scott J. Goetz, and Simon I. Hay. «Fifty Years of Earth Observation Satellites». In: *American Scientist* 96.5 (2008), pp. 390–398. DOI: 10.1511/2008.74.390 (cit. on p. 1).
- [2] Lauren Morone and Carmeyia Gillis. «The History of Numerical Weather Prediction». In: *Mariners Weather Log* 51.3 (2007). URL: https://www.vos.noaa.gov/MWL/dec_07/weatherprediction.shtml (cit. on p. 1).
- [3] EUMETSAT. *Satellites and Instruments – Microwave Imagers and Sounders*. EUMETSAT website. Accessed: April 2026. 2024. URL: <https://www.eumetsat.int/satellites-instruments> (cit. on p. 1).
- [4] EUMETSAT. *Metop-SG Instruments*. EUMETSAT website. Accessed: April 2026. 2022. URL: <https://www.eumetsat.int/metop-sg-instruments> (cit. on p. 2).
- [5] eoPortal. *MetOp-SG (MetOp-Second Generation Program)*. eoPortal website. Accessed: April 2026. 2025. URL: <https://www.eoportal.org/satellite-missions/metop-sg> (cit. on p. 2).
- [6] EUMETSAT. *Metop Second Generation*. EUMETSAT website. Accessed: April 2026. 2020. URL: <https://www.eumetsat.int/metop-sg> (cit. on p. 2).
- [7] ESA. *MetOp-SG satellites*. ESA website. Accessed: April 2026. 2023. URL: https://www.esa.int/Applications/Observing_the_Earth/Meteorological_missions/MetOp_Second_Generation/MetOp-SG_satellites (cit. on p. 2).
- [8] EUMETSAT. *Metop-SG MWI Level 1B Data Guide*. EUMETSAT User Portal. Accessed: April 2026. 2024. URL: <https://user.eumetsat.int/resources/user-guides/metop-sg-mwi-l1b-data-guide> (cit. on pp. 2, 3).
- [9] IEIIT CNR. *MetOp-SG MWI / IEIIT CNR*. IEIIT CNR Website. Accessed: April 2026. 2018. URL: <https://www.ieiit.cnr.it/research/projects/current-projects/metop-sg-mwi> (cit. on p. 3).

- [10] John Xun Yang, Yalei You, William Blackwell, Sidharth Misra, and Rachael A. Kroodsmma. «Quantifying and Characterizing Striping of Microwave Humidity Sounder With Observation and Simulation». In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), p. 5302413. DOI: 10.1109/TGRS.2021.3132560 (cit. on pp. 4, 5, 24, 36).
- [11] Linwei Fan, Fan Zhang, Hui Fan, and Caiming Zhang. «Brief review of image denoising techniques». In: *Visual Computing for Industry, Biomedicine, and Art* 2.1 (2019), p. 7. DOI: 10.1186/s42492-019-0016-7 (cit. on p. 7).
- [12] Zelin Zhang, Hua Li, Yongming Du, Yao Chen, Guoxiang Zhao, Zunjian Bian, Biao Cao, Qing Xiao, and Qinhua Liu. «Stripe Noise Elimination with a Novel Trend Repair Method for Push-Broom Thermal Images». In: *Remote Sensing* 16.17 (2024), p. 3299. DOI: 10.3390/rs16173299 (cit. on pp. 8, 9).
- [13] Marouan Bouali and Saïd Ladjal. «Toward Optimal Destriping of MODIS Data Using a Unidirectional Variational Model». In: *IEEE Transactions on Geoscience and Remote Sensing* 49.8 (2011), pp. 2924–2935. DOI: 10.1109/TGRS.2011.2119399 (cit. on pp. 8, 10, 59, 60).
- [14] Zhengkun Qin, Xiaolei Zou, and Fuzhong Weng. «Analysis of ATMS striping noise from its Earth scene observations». In: *Journal of Geophysical Research: Atmospheres* 118.23 (2013), pp. 13, 214–13, 229. DOI: 10.1002/2013JD020399 (cit. on pp. 9, 10, 24, 46, 48, 59, 62).
- [15] Huijie Dong and Xiaolei Zou. «Striping Noise Mitigation for Tropical Rainfall Measuring Mission Microwave Imager Observations». In: *IEEE Transactions on Geoscience and Remote Sensing* 57.4 (2019), pp. 2449–2459. DOI: 10.1109/TGRS.2018.2873661 (cit. on p. 9).
- [16] Huijie Dong and Xiaolei Zou. «Analysis and Mitigation of Striping Noise in Two Water Vapor Sounding Channels of Global Precipitation Measurement (GPM) Microwave Imager». In: *Journal of Atmospheric and Oceanic Technology* 34.8 (2017), pp. 1693–1707. DOI: 10.1175/JTECH-D-17-0006.1 (cit. on pp. 9, 10, 46–48, 62).
- [17] Yi Chang, Luxin Yan, Tao Wu, and Sheng Zhong. «Remote Sensing Image Stripe Noise Removal: From Image Decomposition Perspective». In: *IEEE Transactions on Geoscience and Remote Sensing* 54.12 (2016), pp. 7018–7031. DOI: 10.1109/TGRS.2016.2594080 (cit. on pp. 9, 10).
- [18] Huanfeng Shen, Wan Jiang, Hongyan Zhang, and Liangpei Zhang. «A piecewise approach to removing the nonlinear and irregular stripes in MODIS data». In: *International Journal of Remote Sensing* 35.1 (2014), pp. 44–53. DOI: 10.1080/2150704X.2013.860564 (cit. on p. 9).

- [19] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. «Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering». In: *IEEE Transactions on Image Processing* 16.8 (2007), pp. 2080–2095. DOI: 10.1109/TIP.2007.901238 (cit. on pp. 10, 50, 60).
- [20] Ymir Mäkinen, Lucio Azzari, and Alessandro Foi. «Collaborative Filtering of Correlated Noise: Exact Transform-Domain Variance for Improved Shrinkage and Patch Matching». In: *IEEE Transactions on Image Processing* 29 (2020), pp. 8339–8354. DOI: 10.1109/TIP.2020.3014721 (cit. on pp. 10, 51).
- [21] Xiaobin Wu, Liangliang Zheng, Chunyu Liu, Tan Gao, Ziyu Zhang, and Biao Yang. «Single-Image Simultaneous Destriping and Denoising: Double Low-Rank Property». In: *Remote Sensing* 15.24 (2023), p. 5710. DOI: 10.3390/rs15245710 (cit. on p. 10).
- [22] Statsmodels Development Team. *Stationarity and detrending (ADF/KPSS)*. Accessed: 2025. 2024. URL: https://www.statsmodels.org/dev/examples/notebooks/generated/stationarity_detrending_adf_kpss.html (cit. on p. 15).
- [23] Lorenzo Galleani and Patrizia Tavella. «The Dynamic Allan Variance». In: *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* 54.3 (2009), pp. 450–464. DOI: 10.1109/TUFFC.2009.1064 (cit. on p. 20).
- [24] The SciPy community. *scipy.signal.welch* — *SciPy Reference Guide*. <https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.welch.html>. Accessed: 2025. 2024 (cit. on p. 21).
- [25] Martti E. Tiuri. «Radio Astronomy Receivers». In: *IEEE Transactions on Antennas and Propagation* 12.7 (1964), pp. 930–938. DOI: 10.1109/TAP.1964.1138345 (cit. on p. 26).
- [26] Alan V. Oppenheim and George C. Verghese. *Wiener Filtering*. Chapter 11, lecture notes for 6.011: Introduction to Communication, Control, and Signal Processing. Accessed: July 7, 2026. 2010. URL: <https://ocw.mit.edu> (cit. on p. 44).
- [27] Zhaohua Wu and Norden E. Huang. «Ensemble empirical mode decomposition: a noise-assisted data analysis method». In: *Advances in Adaptive Data Analysis* 1.1 (2009), pp. 1–41 (cit. on p. 47).
- [28] EUMETSAT. *MWI/ICI L1B Test Data Package (V3): Readme*. Test data package documentation. Accessed: June 2026. EUMETSAT, 2024. URL: <https://user.eumetsat.int/news-events/news/eps-sg-microwave-imager-mwi-and-ice-cloud-imager-ici-l1-b-test-data> (cit. on p. 55).

Acknowledgements

I would like to thank my supervisor, Prof. Enrico Magli, for giving me the opportunity to collaborate with ESA and for his guidance throughout this work.

I am grateful to the Pi School and the Collaborative Innovation Network (CIN) for supporting this project, and to the ESA–ESRIN Φ -Lab for hosting me as a Visiting Researcher. During the three months I spent at ESRIN, I learned about Earth Observation through the work of the many talented people at the Explore Office.

A special thank you goes to the people who worked alongside me on this project: Gabriele Meoni, Lorenzo Papa, Alberto Graziani and Thomas August. Your expertise was essential for me to fully understand the problem, and your guidance and support were invaluable to my academic and personal growth.