Severity and extreme events in sleep apnea

Eduardo Gonçalves
Master in Computational Statistics and Data Analysis
Department of Mathematics
Faculty of Sciences of the University of Porto
2026

Severity and extreme events in sleep apnea

Eduardo Gonçalves

Project Report carried out as part of the Master in Computational Statistics and Data Analysis

Department of Mathematics

2026

Supervisor
Laura Cavalcante, Researcher, Polytechnic of Porto - School of Engineering

Co-supervisor
Margarida Brito, Associate Professor, Faculty of Sciences of Porto University

Acknowledgments

First and foremost, I would like to express my sincere gratitude to both my supervisors, Margarida Brito and Laura Cavalcante, for their guidance, availability, and constructive feedback throughout this work. For all the valuable discussions, the continuous support, and for keeping me focused during the more demanding phases of this journey, your expertise and encouragement were fundamental to the development of this project.

To my friends Hugo, Rafa, André, Nunes, and to those I have not mentioned by name but who are equally important, thank you for the daily exchanges, the adventures we shared, the bad decisions we made together, and all the good moments that made this journey far more enjoyable. You kept pushing me until the end, and that meant more than you know.

To my girlfriend Mariana, thank you for the love and comfort you gave me, especially during the more difficult phases of this work. For your patience and for the belief you always had in me, thank you for being the safe place I can always run to when needed.

To my parents and my brother, thank you for all the moments of peace and unconditional support at home, and for all the chaos in between. You kept my life exciting and never boring, and for that I am truly grateful.

Finally, to my grandparents, and most especially to my grandmother, who, even though she may no longer fully grasp the world around her, still recognizes her grandson and still offers candy. The memories endure in those who come after. This work could not have been completed without the presence of every single one of you, and for all of this, I am deeply grateful.

This work was motivated by previous research on sleep analysis using wearable-device data carried out within the scope of the Remote patient-targeted health monitoring to reduce clinical workload (REMO) project (ITEA-2023-23005- REMO).

Resumo

Este projeto investigou se a duração dos eventos de apneia contém informação clinicamente relevante que o Índice de Apneia-Hipopneia (AHI), por si só, não captura, através da aplicação da Teoria dos Valores Extremos (TVE) à distribuição das durações dos eventos de apneia no conjunto de dados DREAMT. Dos 100 participantes nominais, 90 dispunham de registos de polissonografia (PSG) e de dispositivo vestível utilizáveis, 89 apresentavam pelo menos um evento de apneia registado, e 71 tinham pelo menos dez eventos declustered, permitindo o ajuste individual de um modelo de Distribuição de Pareto Generalizada (GPD) truncada à esquerda ($\tau = 10$ s).

A nível populacional, a forma da cauda da distribuição de durações revelou-se fraca e sensível ao limiar ($\hat{\xi} \approx -0{,}009$), mas a análise individual revelou heterogeneidade substancial entre participantes, aproximadamente 14 % apresentando cauda compatível com comportamento do tipo lei de potência. A aplicação de modelos de Valor Extremo Generalizado (GEV) aos sinais do dispositivo vestível Empatica E4 mostrou que os máximos por época da frequência cardíaca, do intervalo entre batimentos e da temperatura cutânea se enquadram no domínio de Fréchet, enquanto a atividade eletrodérmica e a magnitude do acelerómetro se enquadram no domínio de Weibull.

Foi desenvolvido um classificador combinado de severidade que integra o AHI com a duração máxima individual do evento, o qual reclassificou 24 dos 90 participantes (26,7 %) para uma classe de severidade superior, dos quais 18 (20 %) foram identificados numa zona de risco potencialmente relevante não detetada pelo AHI isolado (AHI < 15, duração máxima > 52 s). Estes resultados foram parcialmente corroborados por comparação com dois conjuntos de dados externos independentes ($n = 1\,000$ e $n = 374$).

Estes resultados sustentam o valor metodológico de incorporar extremos da duração dos eventos na avaliação de severidade da AOS, mas a adoção clínica exigiria validação independente face a desfechos de doentes e a conjuntos de dados PSG prospetivos.

Palavras-Chave: Apneia obstrutiva do sono, Teoria dos Valores Extremos, Distribuição de Valores Extremos Generalizados, classificação de severidade, wearables, DREAMT.

Abstract

This project investigated whether the duration of apnea events carries clinically relevant information that the AHI, on its own, fails to capture, through the application of Extreme Value Theory (EVT) to the distribution of apnea event durations in the DREAMT dataset. Of 100 nominal participants, 90 had usable paired polysomnography (PSG) and wearable recordings, 89 had at least one recorded apnea event, and 71 had at least ten declustered events, enabling individual fitting of a left-truncated Generalized Pareto Distribution (GPD) model ($\tau = 10$ s).

At the population level, the shape of the duration distribution's tail was weak and threshold-sensitive ($\hat{\xi} \approx -0.009$), but individual-level analysis revealed substantial heterogeneity across participants, with approximately 14% displaying a tail compatible with power-law behavior. Applying Generalized Extreme Value (GEV) models to the Empatica E4 wearable signals showed that epoch maxima of heart rate, inter-beat interval, and skin temperature fall within the Fréchet domain, while electrodermal activity and accelerometer magnitude fall within the Weibull domain.

A combined severity classifier integrating the AHI with the maximum individual event duration was developed, upgrading 24 of 90 participants (26.7%) to a higher severity class, of whom 18 (20%) were identified in a potentially clinically relevant zone of risk undetected by the AHI alone (AHI < 15, maximum duration > 52 s). These findings were partially corroborated by comparison against two independent external datasets ($n = 1{,}000$ and $n = 374$).

These findings support the methodological value of incorporating event-duration extremes into OSA severity assessment, but clinical adoption would require independent validation against patient outcomes and prospective PSG datasets.

Keywords: Obstructive sleep apnea, Extreme Value Theory, Generalized Extreme Value Distribution, severity classification, wearables, DREAMT.

Contents

List of Figures

List of Tables

List of Abbreviations

AASM American Academy of Sleep Medicine xi, 1, 8, 9, 16, 20, 21, 31, 37, 53, 58

ACC Accelerometer 11, 26, 27, 28, 29, 47, 56, 58, 60

AHI Apnea-Hypopnea Index ii, iii, vi, viii, xi, 1, 2, 3, 4, 5, 7, 9, 10, 19, 20, 24, 25, 26, 27, 28, 29, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 47, 50, 51, 54, 55, 57, 58, 59, 60, 64, 65, 66, 67, 68, 69

AUC Area Under the Receiver Operating Characteristic Curve 25, 57

BMI Body Mass Index vi, 34, 35, 44, 56, 64

BVP Blood Volume Pulse 11, 26, 27, 29, 45, 57

CCDF Complementary Cumulative Distribution Function vi, 5, 20, 34, 42, 43, 51, 65

CPAP Continuous Positive Airway Pressure 54, 57, 58

DREAMT Dataset for Real-time sleep stage EstimAtion using Multisensor wearable Technology vi, 26, 37, 41, 42, 43, 44, 45, 46, 51, 55, 56, 58

E4 Empatica E4 (wearable device) vii, 26, 45, 47, 48

ECG electrocardiography 8

EDA Electrodermal Activity 11, 26, 27, 28, 29, 45, 47, 51, 56, 57, 58, 59, 60

EEG electroencephalography 8, 11

EMG electromyography 8, 11

EOG electrooculography 8, 11

EVT Extreme Value Theory iii, 2, 3, 4, 5, 6, 7, 10, 12, 13, 19, 20, 26, 41, 43, 45, 53, 56, 58, 59

GEV Generalized Extreme Value Distribution ii, iii, vi, viii, xi, 2, 5, 12, 13, 14, 15, 18, 26, 27, 28, 29, 35, 54, 57, 59, 60

GPD Generalized Pareto Distribution ii, iii, vi, viii, xi, 2, 3, 4, 5, 12, 15, 16, 17, 18, 20, 22, 23, 24, 26, 30, 31, 32, 33, 34, 35, 37, 39, 40, 53, 54, 56, 58, 59, 65

HR Heart Rate 11, 26, 27, 28, 29, 45, 47, 51, 54, 56, 58, 59, 60

IBI Inter-Beat Interval 11, 26, 27, 28, 29, 45, 47, 54, 56, 58, 60

IQR Interquartile Range viii, 33, 36, 41, 64

MLE Maximum Likelihood Estimation 16, 18, 26, 27, 30, 31

NREM Non-Rapid Eye Movement 56, 59

OAHI Obstructive Apnea-Hypopnea Index vi, 34, 35, 64

ODI Oxygen Desaturation Index 47, 48, 57

OSA Obstructive Sleep apnea iii, 1, 2, 4, 5, 7, 8, 9, 10, 11, 25, 28, 29, 33, 35, 41, 42, 43, 44, 45, 50, 52, 54, 55, 56, 57, 59, 60

POT Peaks Over Threshold vi, 2, 4, 5, 15, 16, 20, 21, 29, 30, 43

PPG Photoplethysmography 59

PSG Polysomnography ii, iii, vi, vii, 1, 2, 3, 6, 7, 8, 9, 10, 11, 12, 24, 41, 44, 45, 50, 51, 55, 56, 57, 58, 59, 60

PTAF Pressure Transducer Airflow vii, 11, 45, 51

REM Rapid Eye Movement 7, 51, 56, 59

REMO Remote patient-targeted health monitoring to reduce clinical workload i, 4

ROC Receiver Operating Characteristic 25

SaO2 Arterial Oxygen Saturation vii, 45, 47, 48, 49, 51, 57

SpO2 Peripheral Oxygen Saturation vi, vii, 11, 25, 34, 35, 38, 41, 43, 44, 45, 49, 55, 57, 58, 59, 60

TEMP Skin Temperature 11, 26, 27, 28, 29, 45, 47, 54, 58, 59, 60

List of Notation

ξ	Shape (tail-index) parameter of the GEV/ GPD distribution
σ	Scale parameter (untruncated GPD)
σ_τ	Effective scale parameter above threshold τ
τ	Left-truncation threshold (10 s, AASM minimum)
μ	Location parameter of the GEV distribution
λ_i	Event rate (extreme events per hour) for participant i
$D_i^{\max}$	Maximum observed event duration for participant i
P_{90}, P_{95}, P_{99}	90th/95th/99th percentile of event duration
A_i	AHI-based severity score, $A_i \in \{0, 1, 2, 3\}$
B_i	Duration-based severity score, $B_i \in \{0, 1, 2\}$
C_i	Combined severity score, $C_i = \max(A_i, B_i)$
T_1, T_2	Duration classification thresholds (34 s, 52 s)
x_T	T-hour return level
ρ	Spearman rank correlation coefficient

1. Introduction

1.1. Context and motivation

Obstructive sleep apnea (OSA) is a sleep-related breathing disorder characterized by the repetitive partial or complete collapse of the upper airway during sleep, leading to intermittent cessation of airflow, recurrent hypoxemia, sleep fragmentation, and heightened sympathetic nervous activity (Dempsey et al. 2010). Its epidemiological significance is substantial: it is estimated that approximately one billion adults worldwide are affected by mild to severe OSA, with prevalence exceeding 50% in certain populations and continuing to rise (Benjafield et al. 2019). The consequent health burden encompassing cardiovascular morbidity, cognitive impairment, excessive daytime sleepiness, and reduced quality of life renders effective diagnosis and risk stratification a priority of considerable public health relevance (Qin et al. 2021).

The standard clinical instrument for diagnosing and grading OSA is the Apnea-Hypopnea Index (AHI), computed from overnight polysomnography (PSG) as

$$AHI = \frac{\text{total number of apnea and hypopnea events}}{\text{total sleep time (hours)}}, \tag{1.1}$$

and classified according to the criteria of the American Academy of Sleep Medicine (AASM) into four severity categories: normal (AHI < 5), mild ($5 \leq$ AHI < 15), moderate ($15 \leq$ AHI < 30), and severe (AHI ≥ 30) (Kapur et al. 2017). The AHI has persisted as the dominant severity metric largely because it is simple to calculate and provides an objective event count (Soori et al. 2022). However, a growing body of evidence shows that it incompletely characterizes disease severity and is best regarded as a crude surrogate (Soori et al. 2022; Punjabi 2016).

The fundamental limitation of the Apnea-Hypopnea Index (AHI) is structural: as a frequency-based index, it quantifies *how often* breathing events occur per hour, but contains no information about *how severe* each individual event is. Two dimensions of event severity are systematically ignored. First, the AHI assigns equal weight to apneas and hypopneas despite their markedly

different physiological consequences: apneas, defined by a complete cessation of airflow, are associated with deeper desaturation and greater autonomic perturbation than hypopneas, which involve only a partial reduction in airflow (Soori et al. 2022; Alimardani and Moor 2021). Second, and of central relevance to the present work, the AHI is entirely insensitive to the *duration* of individual events. A patient experiencing few but extremely prolonged apneic episodes may obtain a lower AHI than one with frequent brief events, yet sustain more severe and sustained hypoxic stress. This is not merely a theoretical concern, empirical evidence demonstrates that longer individual event durations produce more severe and prolonged oxygen desaturation events (Kulkas et al. 2017), and that in patients with severe OSA, event duration (not frequency) is associated with increased cardiovascular mortality (Muraja-Murro et al. 2013). Consistent with this, a novel severity metric combining event duration with desaturation depth has been shown to outperform the AHI in predicting cardiovascular risk (Cao et al. 2022). Furthermore, a large historical cohort study of over 13,000 adults with diagnostic PSG found that the composite cardiovascular outcome was predicted by OSA-related factors other than the AHI, including time spent with oxygen saturation below 90%, arousals, and heart rate (Kendzerska et al. 2014). Taken together, these findings establish that the AHI incorporates severity information only to the extent that event severity correlates with event frequency (Soori et al. 2022), an assumption that does not hold in general.

This observation motivates a fundamental statistical question: given that the distribution of individual apnea event durations contains clinically relevant information beyond what the AHI captures, can this distribution, and in particular its tail, be formally characterized in a way that yields interpretable and actionable severity indicators?

Extreme Value Theory (EVT) is the branch of mathematical statistics concerned precisely with the behavior of the tail of a distribution: the probability and magnitude of rare but extreme observations (Coles 2001), with established applications spanning environmental science, finance, and public health (Thomas et al. 2016). Two complementary frameworks are central to its application. The block maxima approach models the distribution of sample maxima through the Generalized Extreme Value (GEV) distribution, whose shape parameter ξ determines whether the tail is bounded, exponential, or polynomial (heavy). The Peaks-Over-Threshold (POT) method, which forms the primary methodological base of this project, models exceedances above a chosen threshold τ through the Generalized Pareto Distribution (GPD):

$$F(x\,;\,\xi,\,\sigma_\tau) \;=\; 1-\left(1+\frac{\xi\,(x-\tau)}{\sigma_\tau}\right)^{-1/\xi}, \quad x>\tau, \tag{1.2}$$

where $\xi \in \mathbb{R}$ is the tail index and $\sigma_\tau > 0$ is the scale parameter above the threshold. When $\xi > 0$, the distribution has a Pareto-type heavy tail, so extreme events remain non-negligibly probable; when $\xi = 0$, the tail decays exponentially; when $\xi < 0$, the distribution has a finite upper endpoint. Applied to the set of individual apnea event durations exceeding a clinically motivated threshold τ (set here at 10 seconds, the minimum clinical definition of an apneic event), the GPD provides a two-parameter characterization of the extreme tail: how heavy it is (ξ) and how spread it is above the threshold (σ_τ). These parameters, which can be estimated individually for each participant, constitute EVT-derived severity indicators that are complementary to the AHI: they describe the intensity of the worst events rather than their frequency.

Beyond the EVT modeling, this work proposes and evaluates a **combined severity classifier** that integrates the AHI with the maximum observed individual event duration into a two-dimensional characterization space. Within this space, a subpopulation emerges that is systematically underclassified by the AHI: participants whose event frequency is low but whose individual event intensity is extreme. This population, designated here as the *undetected risk* group, represents a clinically important failure mode of the AHI that the proposed combined framework is designed to identify.

The empirical foundation of this project is the DREAMT dataset (*Dataset for Real-time sleep stage EstimAtion using Multisensor wearable Technology*) (Wang et al. 2024), comprising 100 overnight PSG-validated recordings from a heterogeneous adult population that includes both healthy volunteers and patients with diverse sleep-disordered breathing conditions. The dataset provides precise temporal annotations of individual apnea, hypopnea, and multiple event episodes, enabling the extraction of event-level duration sequences for each participant, alongside clinical variables including the AHI, mean oxygen saturation, and body mass index. The availability of simultaneously recorded wearable sensor signals (64 Hz) further allows an epoch-level analysis of physiological signal extremes, and a parallel dataset of 100 Hz PSG respiratory signals supports a complementary temporal characterization of oxygen saturation dynamics. To assess generalizability, the findings are compared against two independent publicly available datasets encompassing a combined population of over 1,300 individuals.

This work is carried out within a Master's program at the Universidade do Porto. The methodological emphasis is accordingly placed on the statistical rigor of the EVT framework, including threshold selection, parameter estimation, goodness-of-fit diagnostics, and the interpretation of tail parameters as quantitative severity indicators. The clinical motivation is explicit throughout,

but the primary contribution is methodological: a statistically grounded framework for characterizing apnea severity beyond the AHI.

Beyond this academic setting, this work is connected to the REMO project, a European research initiative, aimed at developing and validating integrated remote patient monitoring solutions that support clinical workflows and reduce clinical workload. The project addresses two clinical use cases: UC1 SPIDA, focused on remote monitoring for spinal disorders rehabilitation, and UC2 MIND-SLEEP, focused on integrated monitoring of sleep and mental well-being for preventive and supportive care. This project is situated within UC2 MIND-SLEEP, where obstructive sleep apnea represents a relevant clinical target. In this context, the severity and temporal pattern of individual respiratory events during sleep constitute clinically relevant information that wearable-based remote monitoring systems may help capture and summarize for clinical teams. The proposed EVT-based severity framework is therefore aligned with the UC2 objective of enabling more nuanced, event-level characterization of sleep-disordered breathing beyond standard frequency-based indices.

The application of EVT to apnea event durations pursued in this project is, to the authors' knowledge, a novel contribution to the intersection of extreme value statistics and clinical sleep medicine: prior work has focused predominantly on event frequency (the AHI) rather than on the statistical distribution of individual event durations, and the tail behavior of this distribution has received no formal statistical treatment. By framing extreme apnea durations as the primary inferential objects and deploying the GPD via the POT method, this project bridges EVT and OSA research, with the goal of constructing a more nuanced severity classifier grounded in the statistical theory of extremes.

1.2. Research questions and objectives

The central research question of this project is:

> *Can Extreme Value Theory, applied to the distribution of individual apnea event durations, provide a statistically sound and clinically interpretable characterization of OSA severity that complements and extends the information contained in the AHI?*

This question is addressed through the following specific objectives:

1. Characterize the empirical distribution of individual apnea event durations and assess evidence of heavy-tail or threshold-sensitive tail behavior through complementary cumulative distribution function (CCDF) analysis and the mean excess plot.

2. Fit GEV distributions to epoch-level maxima of wearable physiological signals and analyze the variation of the tail index ξ across participants and AHI severity categories.

3. Fit a truncated GPD to individual apnea event durations via the POT method, estimate participant-level parameters ξ and σ_τ, and derive theoretical extreme quantiles (P90, P95, P99).

4. Assess the statistical association between EVT-derived parameters and AHI-based severity, and evaluate whether tail parameters carry information about severity independent of event frequency.

5. Develop a duration-based event classifier and a combined severity classifier integrating AHI and maximum event duration, and quantify the population of participants with undetected risk under AHI-only classification.

6. Explore the generalizability of EVT parameters and the combined classifier through cross-dataset comparison against independent external populations.

1.3. Project structure

The remainder of this project is organized as follows.

Chapter 2 provides the theoretical and clinical background necessary to contextualize the methodological contributions. It covers the pathophysiology and epidemiology of OSA, the definition and documented limitations of the AHI, the role of wearable technology in sleep monitoring, and the mathematical foundations of EVT, including the GEV distribution, the GPD, the POT method, threshold selection criteria, and the tail index.

Chapter 3 develops the statistical framework applied to the apnea event duration data. It specifies the left-truncated GPD model and its likelihood, derives the population-level and individual-level estimation procedures, formulates the compound Poisson, GPD risk model, and defines the combined frequency, severity classifier together with its reclassification analysis.

Chapter 4 presents the core methodological and analytical contribution of the project: GEV modeling of epoch-level signal maxima, truncated GPD fitting to apnea event durations with

diagnostic evaluation, analysis of EVT parameters as severity indicators, and the development and evaluation of the duration-based and combined severity classifiers, including the identification of the undetected risk population.

Chapter 5 reports the cross-dataset validation of findings and the supporting evidence from the 100 Hz PSG signal analysis, assessing the generalizability of both the EVT parameters and the combined classifier across independent populations and data modalities.

Chapter 6 synthesizes the contributions of the project, discusses methodological and data limitations, outlines directions for future work, and closes with a synthesis against the six research objectives stated in Section 1.2.

2. Background and related work

This chapter establishes the theoretical and clinical foundations of the present work. Section 2.1 introduces OSA, covering its clinical definition, global prevalence, the diagnostic role of Polysomnography (PSG), and the limitations of the AHI as a severity metric. Section 2.1.3 discusses wearable technology in sleep monitoring and introduces the DREAMT dataset used throughout this project. Section 2.2 develops the statistical framework of Extreme Value Theory (EVT) underlying the methodology applied in later chapters.

2.1. Obstructive Sleep Apnea: clinical overview

2.1.1. Definition, prevalence, and clinical impact

OSA is a chronic sleep-related breathing disorder characterized by repeated episodes of partial or complete obstruction of the upper airway during sleep, resulting in intermittent hypoxemia, hypercapnia, and recurrent cortical arousals (Dempsey et al. 2010). Each obstructive episode is classified as either an *apnea* — a complete cessation of airflow lasting at least ten seconds — or a *hypopnea* — a partial reduction in airflow accompanied by oxygen desaturation or an arousal. The recurrent cycles of airway collapse and re-opening disrupt normal sleep architecture, fragment restorative slow-wave and REM sleep, and impose a sustained burden on the cardiovascular and autonomic nervous systems (Qin et al. 2021).

The global epidemiological burden of OSA is substantial. A landmark meta-analysis by Benjafield et al. estimated that approximately one billion adults aged 30–69 years worldwide suffer from at least mild OSA, with prevalence rates exceeding 50% in some age and sex subgroups (Benjafield et al. 2019). Despite this prevalence, the condition remains severely underdiagnosed: its cardinal symptom — excessive daytime sleepiness — is non-specific, and definitive diagnosis historically requires attended overnight PSG.

The pathophysiology of OSA involves an interplay of anatomical and neuromuscular factors (Dempsey et al. 2010). Craniofacial morphology, adipose deposition in pharyngeal soft tissues, and reduced upper-airway muscle tone during sleep collectively narrow the airway until collapse occurs. The resulting repetitive cycles of obstruction trigger surges in sympathetic nervous system activity, systemic inflammation, and oxidative stress, which explain the strong epidemiological associations between OSA and hypertension, coronary artery disease, atrial fibrillation, stroke, type 2 diabetes, and all-cause mortality (Kendzerska et al. 2014).

2.1.2. Polysomnography, the Apnea-Hypopnea Index, and Its Limitations

Polysomnography as the diagnostic gold standard

In-laboratory PSG is the clinical gold standard for OSA diagnosis (Taran et al. 2021; Seifpour et al. 2018). Figure 2.1 illustrates a representative obstructive apnea event from the DREAMT dataset.

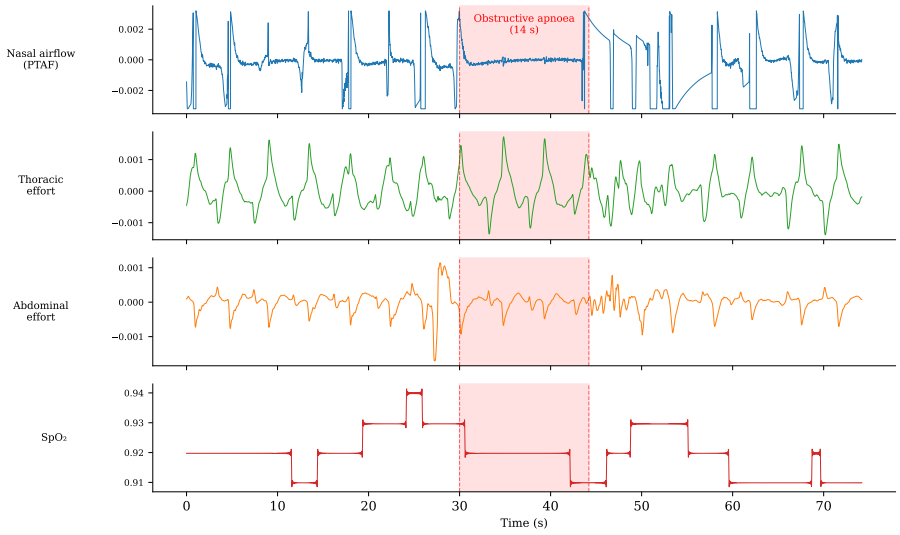

Figure 2.1 - Polysomnography trace - obstructive apnea event (DREAMT dataset, participant S002)

It is a multiparametric overnight recording that simultaneously acquires: electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), electrocardiography (ECG), nasal airflow via thermistor or pressure transducer, thoracic and abdominal respiratory effort, and pulse oximetry (Seifpour et al. 2018). Trained sleep technicians score the recordings epoch by epoch according to the criteria of the American Academy of Sleep Medicine (AASM), which define

discrete sleep stages (N1, N2, N3, and R) and characterize respiratory events with respect to their type, duration, and associated oxygen desaturation or arousal (Kapur et al. 2017).

The Apnea-Hypopnea Index

The primary severity metric derived from PSG is the AHI, defined as the total number of apneas and hypopneas per hour of total sleep time:

$$AHI = \frac{\text{Total apneas} + \text{Total hypopneas}}{\text{Total sleep time (hours)}}. \tag{2.1}$$

The AASM clinical guidelines classify OSA severity on the basis of the AHI into four categories (Kapur et al. 2017):

- **Normal**: AHI < 5 events h^{-1};
- **Mild**: $5 \leq$ AHI < 15;
- **Moderate**: $15 \leq$ AHI < 30;
- **Severe**: AHI ≥ 30.

These thresholds inform clinical decision-making regarding continuous positive airway pressure therapy, mandibular advancement devices, and surgical intervention.

Limitations of the AHI

Despite its wide clinical adoption, the AHI has been subject to substantial methodological criticism (Soori et al. 2022). The following limitations are particularly germane to the motivating questions of this project.

Insensitivity to event duration. The AHI assigns equal weight to each apnea or hypopnea regardless of its duration. A ten-second apnea contributes identically to the index as a ninety-second apnea, despite the substantially greater oxygen desaturation and hemodynamic perturbation associated with the latter (Kulkas et al. 2017). Studies examining individual event durations have demonstrated that longer events produce disproportionate drops in arterial oxygen saturation, suggesting that the distributional properties of event durations — and in particular their extreme values — carry clinically relevant information not captured by the frequency count (Kulkas et al. 2017).

Equivalence of apneas and hypopneas. The AHI aggregates complete obstructions and partial flow reductions into a single count, despite their distinct physiological consequences and differing associations with adverse outcomes (Soori et al. 2022).

Inadequate characterization of hypoxic burden. The AHI does not directly quantify the degree of oxyhemoglobin desaturation associated with respiratory events. Two patients with identical AHI values may experience markedly different nocturnal hypoxemic profiles, and it is the cumulative hypoxic burden — not merely event frequency — that drives many downstream cardiovascular and metabolic sequelae (Punjabi 2016).

Night-to-night and positional variability. The AHI exhibits substantial within-subject variability across nights, leading to documented misclassification of severity when a single diagnostic night is used (Soori et al. 2022). Furthermore, OSA severity is often strongly positional — considerably worse in the supine posture — so the AHI value is confounded by the proportion of the night spent in each position.

Weak discrimination of cardiovascular risk. Prospective cohort studies have shown that, after adjusting for confounders, the AHI is a weaker predictor of incident cardiovascular events than indices of nocturnal hypoxemia and sleep fragmentation (Kendzerska et al. 2014). The severity and duration of individual events appear more directly linked to adverse outcomes than their frequency alone (Muraja-Murro et al. 2013).

These limitations motivate a paradigm shift toward event-level analysis and, specifically, toward characterizing the statistical distribution of individual event durations. The present work applies EVT to this distributional problem, treating extreme apnea durations as the primary objects of inferential interest.

2.1.3. Wearable technology in sleep monitoring

The logistical and economic constraints of in-laboratory PSG — high cost, limited scalability, first-night effect, and the artificial sleeping environment — have driven increasing interest in ambulatory and wearable alternatives (Wang et al. 2024). Consumer-grade and research-grade wearable devices now permit continuous, multi-night, home-based monitoring of physiological signals relevant to sleep-disordered breathing.

The **DREAMT** (Dataset for Real-time sleep stage EstimAtion using Multisensor wearable Technology) dataset, introduced by Wang et al., is the primary data source for this project (Wang et al. 2024). It comprises simultaneous PSG recordings at 100 Hz and wrist-worn Empatica E4 recordings at 64 Hz from 100 participants, including OSA patients and healthy controls.

Of the 100 nominal participants in the released DREAMT dataset, ten lacked the corresponding raw signal files and were excluded outright, yielding $n = 90$ participants with usable paired PSG and wearable recordings. A further participant remained awake for the entire recording and consequently exhibited no scoreable apnea events; this physiological outlier was retained for the epoch-level wearable-signal analyses of Chapter 4 but excluded from the event-duration analyses, which are therefore based on a cohort of $n = 89$ participants with at least one recorded apnea event. A subsequent restriction to participants with a sufficient number of declustered events, applied for the robustness of individual-level estimation, is described in Section 3.3 and yields the final analytic cohort of $n = 71$ used throughout Sections 3.3, 3.5 and 4.4. Unless otherwise stated, all other analyses in this project (including the epoch-level signal extremes of Chapter 4) use the $n = 90$ cohort.

The Empatica E4 wearable captures:

- **Photoplethysmography (BVP):** from which HR and Inter-Beat Interval (IBI) may be derived;
- **Electrodermal Activity (EDA):** reflecting sympathetic arousal associated with apneic events;
- **Skin temperature (TEMP):** a marker of peripheral vasomotor tone;
- **Three-axis accelerometry (ACC):** enabling posture estimation and motion artifact detection.

Figure 2.2 summarizes this simultaneous recording setup.

The PSG channels include nasal airflow (PTAF), thoracic and abdominal respiratory effort, and pulse oximetry (SpO2), together with EEG, EOG, and EMG. The dataset provides per-event annotations of apnea type (obstructive, central, mixed), onset time, and duration, enabling the event-level statistical analysis pursued here.

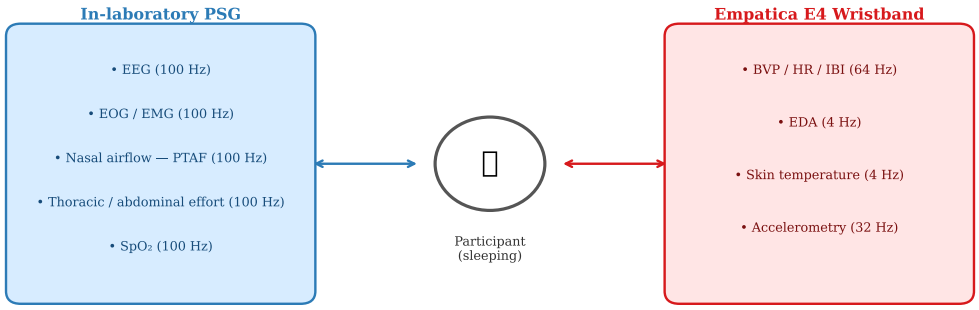

Figure 2.2 - DREAMT dataset - simultaneous PSG + wearable recording

Wearable monitoring offers the advantages of ecological validity, longitudinal data collection, and scalability, which are particularly relevant for studying disease progression and evaluating treatment response across multiple nights. Limitations include susceptibility to motion artifacts, reduced signal fidelity relative to clinical-grade sensors, and the absence of direct cortical arousal scoring (Wang et al. 2024). Beyond sleep-specific applications, wearable-derived physiological monitoring has recently been integrated into broader remote-health ecosystems, including AI-supported elderly monitoring and medication adherence platforms (Vieira et al. 2026) and workflow-oriented remote patient monitoring initiatives designed to reduce clinical workload (Lima et al. 2026), underscoring the growing clinical relevance of the wearable-signal analyses pursued in this project.

2.2. Extreme Value Theory: theoretical foundations

EVT is the branch of probability theory and statistics concerned with the asymptotic behavior of extreme phenomena — events in the far tail of a distribution, frequently beyond the range of previously observed data. Originating in the 1920s and consolidated through the seminal contributions of Fisher, Tippett, and Gnedenko, EVT provides a rigorous framework for modeling distributional tails and for extrapolating to quantile levels beyond direct observation (Coles 2001). The foundational text for the statistical methodology employed in this project is (Coles 2001), which presents the GEV distribution, the GPD, threshold selection, return level estimation, and model diagnostics within a unified likelihood-based framework.

2.2.1. Block Maxima and the Generalized Extreme Value Distribution

The Extremal Types Theorem

The classical approach to extreme value modeling proceeds from the *block maxima* method. Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (i.i.d.) random variables with common distribution function F. Define the block maximum over n observations as

$$M_n = \max\{X_1, \ldots, X_n\}. \tag{2.2}$$

The exact distribution of M_n is $[F(z)]^n$, which is intractable in practice since F is unknown. The fundamental result of EVT states that if there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$\Pr\left\{\frac{M_n - b_n}{a_n} \leq z\right\} \rightarrow G(z) \quad \text{as } n \rightarrow \infty, \tag{2.3}$$

for a non-degenerate distribution function G, then G must belong to the **Generalized Extreme Value Distribution (GEV)** family (Coles 2001):

$$G(z;\, \mu, \sigma, \xi) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, \tag{2.4}$$

defined on $\{z : 1 + \xi(z-\mu)/\sigma > 0\}$, where $\mu \in \mathbb{R}$ is the *location* parameter, $\sigma > 0$ is the *scale* parameter, and $\xi \in \mathbb{R}$ is the *shape* parameter. The GEV family unifies three classical sub-families according to the sign of ξ (Coles 2001):

- **Gumbel** ($\xi = 0$): interpreted as the limit of (2.4) as $\xi \rightarrow 0$, yielding $G(z) = \exp\{-e^{-(z-\mu)/\sigma}\}$. Applicable to light-tailed parent distributions (e.g., Normal, Gamma).
- **Fréchet** ($\xi > 0$): characterized by a polynomial (heavy) tail with index $1/\xi$. Applicable to power-law distributions; arises naturally for financial returns, environmental extremes, and — as investigated in this project — apnea event durations.
- **Weibull** ($\xi < 0$): characterized by a finite upper endpoint at $\mu - \sigma/\xi$. Applicable to bounded distributions.

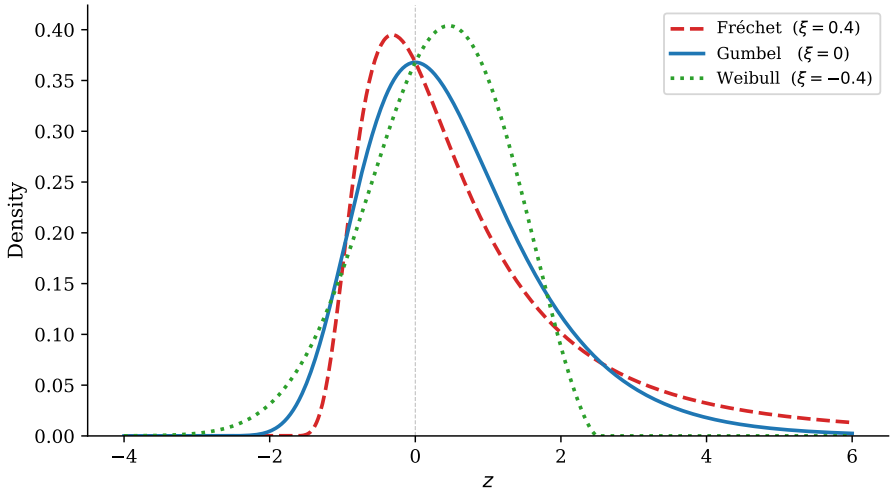

Figure 2.3 - GEV probability density functions ($\mu = 0, \sigma = 1$ varying ξ)

Return levels

A key output of GEV modeling is the *return level* z_p, defined as the level exceeded by the block maximum in any given block with probability p: $G(z_p) = 1 - p$. Inverting (2.4) yields

$$z_p = \mu - \frac{\sigma}{\xi}\left[1 - \{-\log(1-p)\}^{-\xi}\right], \quad \xi \neq 0. \tag{2.5}$$

The *return period* $1/p$ represents the expected number of blocks between successive exceedances of z_p. Return level estimates and their confidence intervals, obtained via the delta method or profile likelihood (Coles 2001), are the primary inferential outputs of block maxima analyses.

Inference

Parameters (μ, σ, ξ) are typically estimated by maximum likelihood. The log-likelihood for m independent block maxima $z_1, \ldots, z_m$ under the GEV model ($\xi \neq 0$) is (Coles 2001):

$$\ell(\mu, \sigma, \xi) = -m \log \sigma - \left(1 + \tfrac{1}{\xi}\right) \sum_{i=1}^{m} \log\left[1 + \xi\left(\tfrac{z_i - \mu}{\sigma}\right)\right] - \sum_{i=1}^{m} \left[1 + \xi\left(\tfrac{z_i - \mu}{\sigma}\right)\right]^{-1/\xi}, \tag{2.6}$$

subject to $1 + \xi(z_i - \mu)/\sigma > 0$ for all i. Goodness of fit is subsequently assessed by comparing empirical and model-based probabilities, quantiles, and return levels graphically, via the probability plot, quantile plot, and return level plot, respectively (Coles 2001).

2.2.2. Peaks-Over-Threshold and the Generalized Pareto Distribution

Motivation for the POT approach

The block maxima method discards potentially valuable extreme observations that are not the block maximum, resulting in information loss. The *Peaks Over Threshold (POT)* method addresses this deficiency by modeling *all* observations exceeding a sufficiently high threshold u, retaining more of the tail data and avoiding arbitrary choices of block length (Coles 2001).

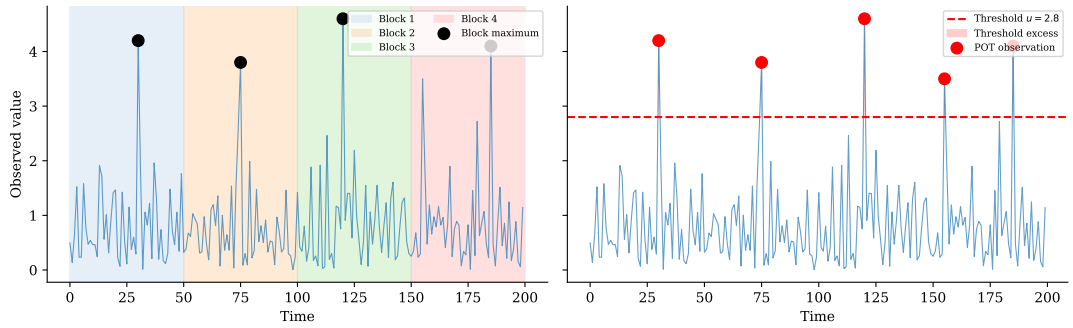

Figure 2.4 - Comparison of extreme value modeling approaches

The Pickands-Balkema-de Haan theorem

Let $Y = X - u \mid X > u$ denote the excess above threshold u. The **Pickands-Balkema-de Haan theorem** states that if the distribution of block maxima converges to a GEV distribution with shape parameter ξ, then for a sufficiently large threshold u, the conditional distribution of excesses is approximately the **Generalized Pareto Distribution (GPD)** (Coles 2001):

$$H(y;\, \sigma_u, \xi) = 1 - \left(1 + \frac{\xi\, y}{\sigma_u}\right)^{-1/\xi}, \tag{2.7}$$

for $y > 0$ when $\xi \geq 0$, or $0 < y < -\sigma_u/\xi$ when $\xi < 0$, where $\sigma_u = \sigma + \xi(u - \mu) > 0$ is the threshold-reparameterized scale parameter. Crucially, the shape parameter ξ in (2.7) is identical to that in the parent GEV distribution (2.4), providing a direct theoretical link between the two frameworks (Coles 2001).

Special cases of the GPD include:

- $\xi = 0$: the exponential distribution $H(y) = 1 - e^{-y/\sigma_u}$, corresponding to a light (exponential) tail;

- $\xi > 0$: a Pareto-type distribution with polynomial tail $\bar{H}(y) \sim y^{-1/\xi}$ — heavy-tailed behavior;
- $\xi < 0$: a bounded distribution with finite upper endpoint at $-\sigma_u/\xi$.

Weissman extrapolation

Once the GPD is fitted by MLE to the threshold excesses, extreme quantiles beyond the range of the data may be estimated via the *Weissman-type* extrapolation formula (Coles 2001; Rutikanga, Diop, and Uwilingiyimana 2024):

$$\hat{q}_{1-p} = u + \frac{\hat{\sigma}_u}{\hat{\xi}} \left[\left(\frac{\hat{p}_u}{p} \right)^{\hat{\xi}} - 1 \right], \tag{2.8}$$

where $\hat{p}_u = k/n$ is the estimated probability of exceeding u (with k exceedances in n observations) and $p < \hat{p}_u$ is the target exceedance probability. This formula underpins the extreme quantile estimation framework used in Chapter 3 of this project, in other words, it extrapolates from the observed exceedance rate at u to the (typically unobserved) quantile associated with a smaller target probability p, using the fitted GPD shape ξ to control the rate of extrapolation: a larger ξ implies a slower-decaying tail and therefore a more aggressive (larger) extrapolated quantile for a given p.

The POT framework is particularly well-suited to the analysis of apnea event durations. Apnea events form a sequence of non-negative random variables for which the clinical interest lies precisely in the tail of the duration distribution: events of extreme duration drive the most severe hemodynamic perturbations (Kulkas et al. 2017). Modeling excesses over a duration threshold — corresponding, for instance, to the AASM minimum apnea criterion of ten seconds — aligns the statistical framework directly with the clinical question.

2.2.3. Threshold selection and model diagnostics

A critical component of POT analysis is the selection of the threshold u. The threshold must be chosen high enough to justify the GPD approximation, yet low enough to retain sufficient exceedances for reliable parameter estimation, a classical bias-variance trade-off (Coles 2001).

The mean excess function

The **mean excess function** (also called the mean residual life function) of a random variable X is defined as

$$e(u) = \mathbb{E}[X - u \mid X > u], \tag{2.9}$$

the expected excess above threshold u, conditional on exceedance. A fundamental property of the GPD is that $e(u)$ is linear in u above any valid threshold (Coles 2001):

$$e(u) = \frac{\sigma_0 + \xi\, u}{1 - \xi}, \quad \xi < 1, \tag{2.10}$$

where σ_0 is the GPD scale parameter at a reference threshold. The **sample mean excess plot** (or *meplot*) displays the empirical estimate using the indicator function $\mathbf{1}\{\cdot\}$ (equal to 1 if the condition holds and 0 otherwise),

$$\hat{e}(u) = \frac{1}{k(u)} \sum_{i=1}^{n} (X_i - u)\, \mathbf{1}\{X_i > u\}$$

against u, where $k(u)$ is the number of observations exceeding u.

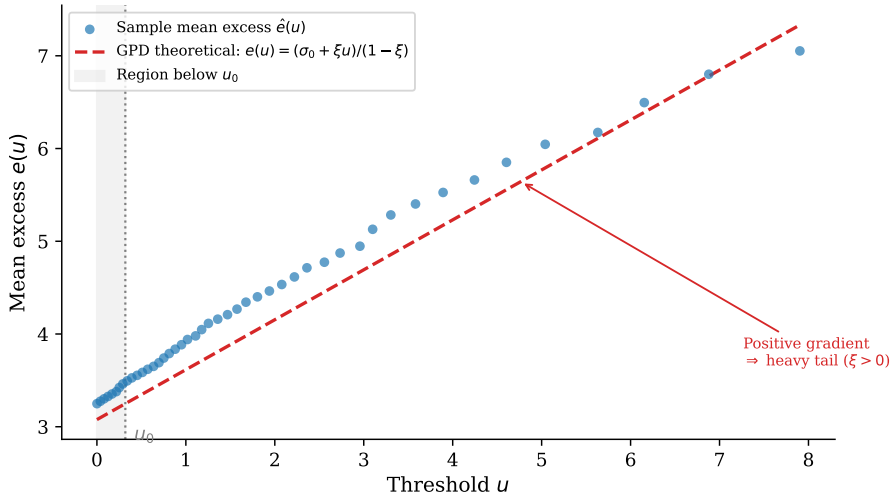

Figure 2.5 - Mean excess plot (simulated GPD, $\xi = 0.35, \sigma = 2$)

Threshold selection proceeds by identifying the lowest value of u above which the plot is approximately linear. A *positive gradient* in the mean excess plot above the chosen threshold is diagnostic of a heavy-tailed distribution with $\xi > 0$ (Coles 2001; Rutikanga, Diop, and Uwilingiyimana 2024), as illustrated in Figure 2.5 for a simulated heavy-tailed GPD sample.

Parameter stability plots

Complementary to the mean excess plot, **parameter stability plots** display the MLE estimates of the reparameterized scale $\sigma^* = \sigma_u - \xi\, u$ (which is invariant to threshold choice under the GPD model) and the shape parameter ξ as functions of u. Under the correct model, both quantities should be approximately constant above the true threshold (Coles 2001).

2.2.4. Tail index and heavy-tail behavior

The shape parameter ξ of the GEV and GPD governs the qualitative character of tail behavior and is referred to in the extreme value literature as the *tail index* or *extreme value index* (Coles 2001; Rutikanga, Diop, and Uwilingiyimana 2024).

Domains of attraction

The value of ξ determines the *domain of attraction* of the parent distribution F: the class of distributions whose block maxima, after suitable normalization, converge to the same limiting GEV shape. Three domains are distinguished according to the sign of ξ:

- **Fréchet domain** ($\xi > 0$): the distributional tail decays polynomially, $\bar{F}(x) \sim x^{-1/\xi}$ as $x \to \infty$. The distribution is *heavy-tailed* or *regularly varying with index* $-1/\xi$. Moments of order $k \geq 1/\xi$ do not exist. Canonical examples include the Pareto, log-Gamma, and Burr distributions.
- **Gumbel domain** ($\xi = 0$): the tail decays faster than any power law (e.g., exponentially or faster), all moments are finite, and $e(u)$ is approximately constant above the threshold.
- **Weibull domain** ($\xi < 0$): the distribution has a finite right endpoint $x^* = \mu - \sigma/\xi$ beyond which no observations are possible.

Implications for apnea event durations

The empirical observation of occasional extreme apnea events lasting 120 seconds or more, substantially in excess of the bulk of the distribution, suggests heavy-tailed behavior ($\xi > 0$) in the duration distribution. Confirming this through formal EVT inference, and estimating ξ by AHI severity class and apnea type, constitutes a central objective of this project. If $\xi > 0$, the return level curve (2.5) is concave and unbounded, implying a non-negligible probability of extreme events of very large duration that would be severely underestimated by frequency-based metrics such as the AHI.

3. Statistical methods for extreme apnea duration analysis

This chapter develops the statistical framework applied to the apnea event duration data. Section 3.1 introduces the relevant framework from EVT. Section 3.2 specifies the left-truncated GPD model and derives the likelihood used for estimation. Section 3.3 describes the individual-level GPD fitting procedure and the participant-level summary statistics derived from it. Section 3.4 formulates the compound Poisson- GPD risk model and derives participant-level return levels. Section 3.5 defines the combined frequency-severity classifier and evaluates its discriminative performance against the standard AHI-alone classification.

3.1. Choice of the POT framework and preliminary tail diagnostics

The theoretical foundations of the Peaks-Over-Threshold (POT) approach and the Generalized Pareto Distribution (GPD), the Pickands-Balkema-de Haan theorem, the GPD density, the domains of attraction, and the mean excess function, are presented in Section 2.2.2. For the purposes of this chapter, the POT approach is adopted because it: (1) makes full use of all scored respiratory events rather than only one per block period; (2) has a natural, clinically interpretable threshold in $\tau = 10$ s, the AASM minimum scored duration (Kapur et al. 2017); and (3) yields closed-form expressions for return levels directly interpretable in terms of event frequency and duration severity.

Figure 3.1 presents the four standard heavy-tail diagnostics (log-log histogram, log-log CCDF, mean excess plot, and Hill plot) computed on the pooled 8 931 raw event durations (before declustering). The Hill plot is based on the Hill estimator of the tail index (Hill 1975): given order statistics $X_{(1)} \leq \cdots \leq X_{(n)}$, it is defined as

$$H_k = \frac{1}{k} \sum_{i=1}^{k} \Big(\log X_{(n-i+1)} - \log X_{(n-k)} \Big), \tag{3.1}$$

where $k < n$ is the number of upper-order statistics used, under a Pareto-type tail, $H_k \xrightarrow{p} \xi$ as $k, n \to \infty$ with $k/n \to 0$; the Hill *plot* displays H_k against k and a stable plateau indicates a well-defined tail index. The statistical properties of H_k in the context of tail prepivoting are studied in (Brito, Freitas, and Freitas 2016). These diagnostics are broadly consistent with heavy-tailed behavior at low thresholds; however, this visual, exploratory evidence attenuates as the threshold is raised, as shown by the near-zero population-level shape estimate and the instability of $\hat{\xi}$ across thresholds (Section 4.2).

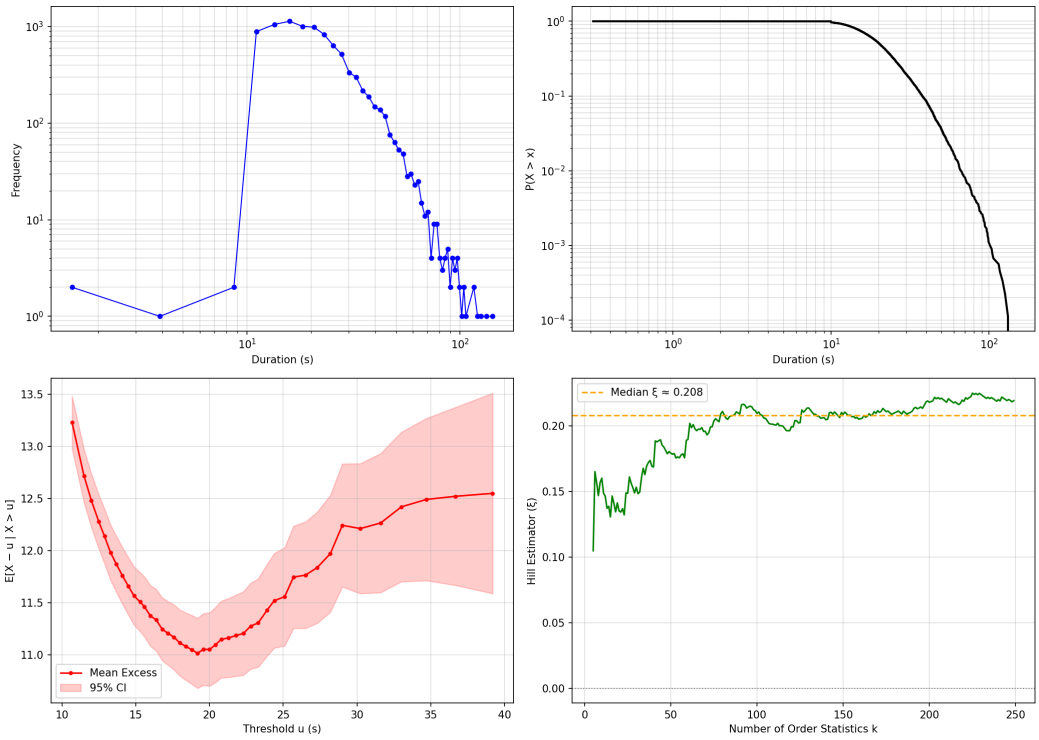

Figure 3.1 - Heavy-tail evidence in event durations

3.2. Left-truncated GPD model

3.2.1. Motivation for left truncation

The AASM scoring rules require that a respiratory event be at least 10 s in duration to be counted (Kapur et al. 2017). Let $d_1, \ldots, d_n$ denote the observed apnea event durations (pooled across participants at the population level, or for a single participant at the individual level). Consequently, the observed data d_k are subject to a fixed left-truncation at $\tau = 10s$: only events with $d_k \geq \tau$ enter the dataset. Standard POT methodology models the excesses $Y_k = d_k - \tau$ for

events exceeding a chosen analysis threshold τ, treating τ as a free modeling choice. Here, however, $\tau = 10s$ is not a free threshold but a hard scoring floor: by construction, no event below 10 s exists in the data. Fitting the standard (untruncated) GPD to the raw excesses $d_k - \tau$ would ignore this truncation mechanism, the fact that the entire sample is already conditioned on $d_k \geq \tau$, leading to biased parameter estimates (Grimshaw 1993). The left-truncated GPD adopted here conditions the entire likelihood on the event $X \geq \tau$, explicitly conditioning on the clinical threshold $\tau = 10$ s and providing an interpretable parametrization for the observable distribution of events. The scale parameter that governs the observable distribution is the *effective scale*:

$$\sigma_\tau = \sigma + \xi\,\tau, \tag{3.2}$$

where σ is the underlying (untruncated) scale.

3.2.2. Likelihood and declustering

Before fitting, temporal dependence between consecutive events is addressed by *declustering*: events separated by a gap of less than 60 s are grouped into a cluster, and only the maximum-duration event within each cluster is retained. This procedure reduces the 8 931 raw events to 2 262 declustered events (a 74.7 % reduction), yielding an approximately independent sample to which likelihood inference can be validly applied. The 60-second gap threshold is motivated by the typical duration of inter-event recovery periods observed in the cohort and is conservative relative to standard climatological practice (Ferro and Segers 2003).

Let $\mathbf{d} = (d_1, \ldots, d_n)$ denote the resulting vector of declustered durations, with $d_k \geq \tau$ for all k (cf. Section 3.2.1). Under the GPD(ξ, σ) with left-truncation at τ, the contribution of each observation to the log-likelihood is the log of the conditional density:

$$\log f(d_k \mid d_k \geq \tau; \xi, \sigma) = \log f_{\xi,\sigma}(d_k) - \log S_{\xi,\sigma}(\tau), \tag{3.3}$$

where $f_{\xi,\sigma}$ and $S_{\xi,\sigma}$ are the GPD density and survival function, respectively. The full log-likelihood is:

$$\ell(\xi, \sigma; \mathbf{d}) = -n \log \sigma - \left(\frac{1}{\xi} + 1\right) \sum_{k=1}^{n} \log\left(1 + \frac{\xi d_k}{\sigma}\right) + \frac{n}{\xi} \log\left(1 + \frac{\xi\tau}{\sigma}\right), \qquad \xi \neq 0. \tag{3.4}$$

Maximum likelihood estimates $(\hat{\xi},\ \hat{\sigma})$ are obtained numerically by minimizing the negative log-likelihood (3.3) using L-BFGS-B (Limited-memory Broyden–Fletcher–Goldfarb–Shanno with box constraints), a quasi-Newton optimization algorithm well suited to bounded parameter estimation,

with multiple starting points and bounds $\xi \in (-0.49, 2.5)$, $\sigma \in (0.1, 10 \cdot d_{\max})$, following the recommendations of (Coles 2001). A fit is deemed to have converged when the achieved negative log-likelihood is finite and the bound constraints are not active.

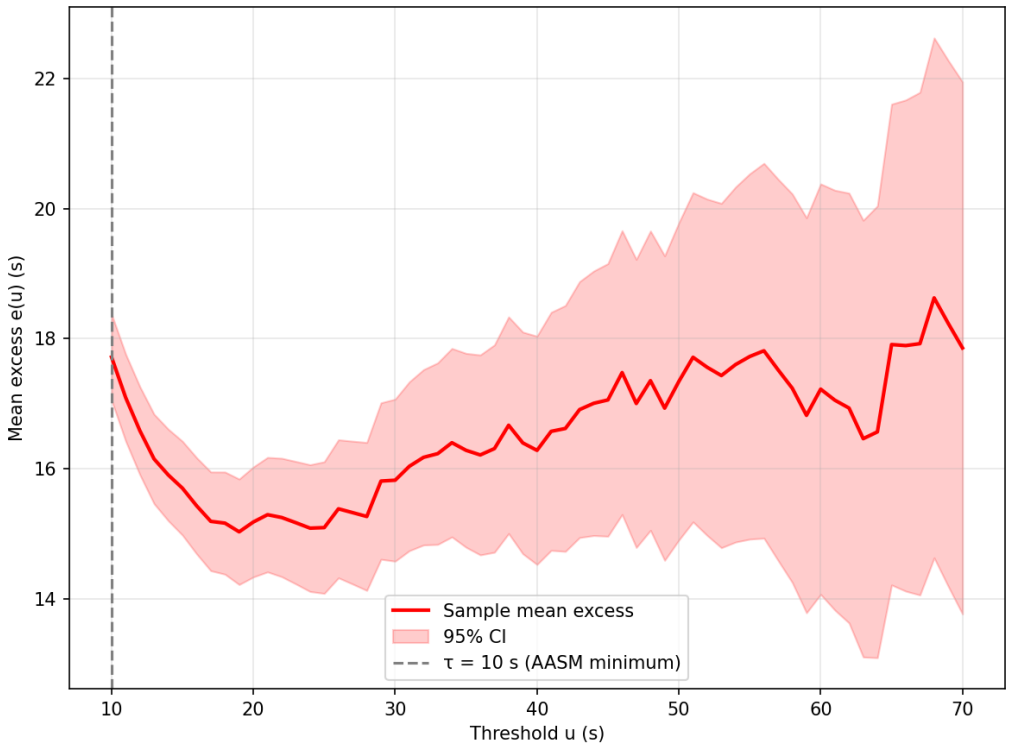

Figure 3.2 - Mean excess plot (pooled declustered apnea durations, n = 2262)

3.3. Individual GPD parameter estimates

The left-truncated GPD is fitted separately for each of the 90 participants in the dataset, of whom 71 had at least ten declustered events (N $\geq$ 10) to enable individual GPD fitting, conditioning the likelihood on the fixed clinical threshold $\tau = 10$ s as in Section 3.2 (not a data-driven percentile threshold). Individual estimates $(\hat{\xi}_i, \hat{\sigma}_i)$ and the corresponding effective scale $\hat{\sigma}_{\tau,i} = \hat{\sigma}_i + \hat{\xi}_i \cdot \tau$ are stored alongside the participant-level summary statistics:

- $D_i^{\max}$ — the duration of the single longest apnea event recorded for participant i;
- $P_{95,i}$ — empirical 95th percentile of individual event durations;
- λ_i — event rate (events per hour of total sleep time), estimated as $\lambda_i = \left(N_i^{\text{declust}} / N_i^{\text{total}}\right) AHI_i$;

- A_i — AHI severity score (equation (2.1)).

3.4. Compound Poisson-GPD risk model

The compound Poisson- GPD model combines two components:

Frequency component. The occurrence of extreme apnea events above τ is modeled as a homogeneous Poisson process with intensity λ_i (events per hour of sleep), estimated individually as described in Section 3.3.

Severity component. Conditional on an event occurring, its duration follows the truncated GPD$(\hat{\xi}_i, \hat{\sigma}_i)$ fitted to participant i's declustered exceedances.

The T-hour return level x_T is defined as the duration exceeded on average at most once per T hours of sleep:

$$x_T = \tau + \frac{\hat{\sigma}_{\tau,i}}{\hat{\xi}_i}\left[(\lambda_i T)^{\hat{\xi}_i} - 1\right], \qquad \hat{\xi}_i \neq 0. \tag{3.5}$$

For a typical overnight PSG session of $T = 8$ h, this equation yields a clinically interpretable worst-expected-event duration for each participant under the assumed stationarity of the Poisson process. The empirical return level curves, the risk-quadrant analysis, and their clinical interpretation are presented in Section 4.5.4 (Figure 4.12).

3.5. Combined frequency-severity classifier

3.5.1. Definition

The combined severity classifier assigns each participant i to one of four ordinal classes by taking the element-wise maximum of the AHI severity score $A_i \in \{0, 1, 2, 3\}$ and a duration-based score $B_i \in \{0, 1, 2\}$:

$$C_i = \max(A_i,\ B_i)\,, \qquad B_i = \begin{cases} 0 & D_i^{\max} \leq T_1, \\ 1 & T_1 < D_i^{\max} \leq T_2, \\ 2 & D_i^{\max} > T_2, \end{cases} \tag{3.6}$$

where T_1 and T_2 are duration thresholds and the combined score C_i maps back to the four-class ordinal label {Normal, Mild, Moderate, Severe} as defined in Section 2.1.2. By design, the combined classifier *never downgrades* a participant: $C_i \geq A_i$ for all i.

3.5.2. Threshold selection

The reference threshold pair $(T_1, T_2) = (34\text{ s}, 52\text{ s})$ is grounded in clinical evidence linking event duration to SpO2 nadir: events of 34 s are associated with $\text{SpO}_2 \leq 85\,\%$, and events of 52 s with $\text{SpO}_2 \leq 80\,\%$ (Zhou et al. 2026).

To assess whether data-driven thresholds improve discriminative performance, a grid search is conducted over all pairs (T_1, T_2) with $T_1 \in [5, 90]$ s, $T_2 \in [10, 180]$ s, $T_2 > T_1 + 4$ s (11 051 combinations), evaluating four metrics:

1. Spearman ρ between B_i and continuous AHI (primary criterion);
2. linear weighted κ between B_i and A_i;
3. AUC- ROC for AHI ≥ 15 (Moderate+Severe vs Normal+Mild);
4. AUC- ROC for AHI ≥ 30 (Severe vs rest).

Table 3.1 compares the reference and optimal pairs.

Table 3.1 - Performance of the duration-based score B_i at the clinical reference thresholds and at the data-optimal pair (maximum Spearman ρ) across four evaluation metrics.

Pair	T_1	T_2	**Spearman** ρ	κ	**AUC**$_{\geq 15}$	**AUC**$_{\geq 30}$
Reference	34 s	52 s	0.407	0.253	0.685	0.582
Data-optimal	21 s	39 s	**0.546**	**0.328**	0.716	0.664
AUC-optimal	39 s	78 s	0.499	0.254	**0.742**	**0.696**

The data-optimal pair $(21\text{ s}, 39\text{ s})$ achieves the highest Spearman correlation ($\rho = 0.546$, versus 0.407 for the reference), while the AUC-optimal pair $(39\text{ s}, 78\text{ s})$ maximizes discrimination of moderate and severe OSA (AUC $= 0.742$, versus 0.685 for the reference). The clinical reference pair is retained as the primary specification throughout this project because its thresholds have direct physiological interpretation linked to SpO2 desaturation severity; the data-driven alternatives are reported in Table 3.1 as a sensitivity check.

4. Extreme value modeling and duration-based risk classification

This chapter presents the empirical results of applying the EVT framework of Chapter 3 to the DREAMT cohort. Section 4.1 analyzes the tail behavior of the six wearable physiological signals at the epoch level using the GEV distribution. Section 4.2 presents the population-level GPD fit to the pooled declustered sample and examines threshold sensitivity of the shape estimate. Section 4.3 reports the truncated GPD fits for apnea event durations and derives participant-level extreme quantiles. Section 4.4 examines the statistical associations between the fitted EVT parameters and AHI severity. Section 4.5 presents the duration-based classifier, the two-dimensional risk space, and the combined frequency-severity reclassification. The GEV analysis of wearable signals is presented first as contextual characterization of EVT behavior at the physiological signal level; the GPD analysis of apnea event durations in Sections 4.2-4.5 constitutes the primary analytical contribution.

4.1. GEV modeling of wearable signal epoch maxima

4.1.1. Pooled fitting and signal parameter estimates

For each participant and each of the six E4 signals, HR, IBI, EDA, TEMP, BVP, and Accelerometer (ACC) magnitude, the epoch-level block maximum (one per 30 s epoch) was extracted and a GEV distribution was fitted by MLE. The full pooled analysis uses 7 275 epoch maxima aggregated across the cohort (Table 4.1).

The signals split into two groups by the sign of $\hat{\xi}$. EDA and ACC magnitude exhibit $\hat{\xi} \approx -1$, consistent with a Weibull distribution with a finite upper bound, reflecting the physiological saturation of these sensors. HR, IBI, and TEMP exhibit positive pooled shape parameters, formally placing their epoch maxima in the Fréchet domain, but the strength of this evidence differs substantially

Table 4.1 - Pooled GEV parameter estimates for the six wearable signals.

Signal	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	**Domain**
HR (bpm)	65.5	11.5	0.027	Fréchet (marginal)
IBI (s)	1.04	0.22	0.153	Fréchet
TEMP (°C)	33.6	2.16	0.543	Fréchet
EDA (μS)	0.31	0.33	-0.964	Weibull
ACC_{mag} (mg)	67.2	5.6	-0.994	Weibull
BVP (a.u.)	—	—	-5.12	Weibull

Note: $\hat{\mu}$ and $\hat{\sigma}$ are omitted for BVP because the MLE fit did not converge to numerically stable values; the extreme estimate $\hat{\xi} = -5.12$ is indicative of sensor saturation rather than a physiologically meaningful heavy or bounded tail. The HR estimate ($\hat{\xi} = 0.027$) is close to the Gumbel boundary and is reported as marginal evidence of heavy-tailedness, not a robust Fréchet-domain finding (Section 4.1.1).

across the three signals. TEMP shows a positive estimate ($\hat{\xi} = 0.543$), which remains stable and consistently positive across all four AHI severity classes ($\hat{\xi} \in [0.40, 0.65]$, Section 4.1.2), supporting a genuinely heavy-tailed interpretation. IBI shows a more modest but still clearly positive estimate ($\hat{\xi} = 0.153$). HR, by contrast, yields a pooled $\hat{\xi} = 0.027$, a value of the same small magnitude as the population-level apnea duration estimate judged indistinguishable from zero in Section 4.2; consistent with this, the severity-stratified estimates for HR change sign across classes (Mild: 0.086; Moderate: -0.013; Severe: -0.137, Section 4.1.2), indicating that the pooled positive estimate is not a stable Fréchet-domain finding but reflects sampling variability around a boundary value close to zero. BVP yields an extreme negative estimate ($\hat{\xi} = -5.12$) that reflects a numerically bounded distribution rather than a physiologically meaningful heavy tail, and is treated with caution in subsequent analyses.

Figure 4.1 shows the Q-Q plots of the pooled epoch maxima against the fitted GEV quantiles for each signal. Alignment along the 1:1 reference line is good for TEMP and reasonable for IBI, supporting the adequacy of the GEV approximation for these two signals; the HR quantile plot shows visible curvature relative to the 1:1 line, consistent with the marginal, threshold-sensitive shape estimate discussed above (Section 4.1.1) rather than a robust Fréchet-domain fit. Systematic departure in the EDA and ACC tails reflects the bounded nature of those distributions.

4.1.2. Tail index by AHI severity: statistical comparison

Table 4.2 stratifies the pooled GEV shape estimates by AHI severity group, fitting one GEV per signal–severity combination. The epoch counts differ across groups owing to the unequal numbers of participants per severity class and, more substantially, to variability in effective recording length: the per-participant mean epoch count decreases monotonically from Normal

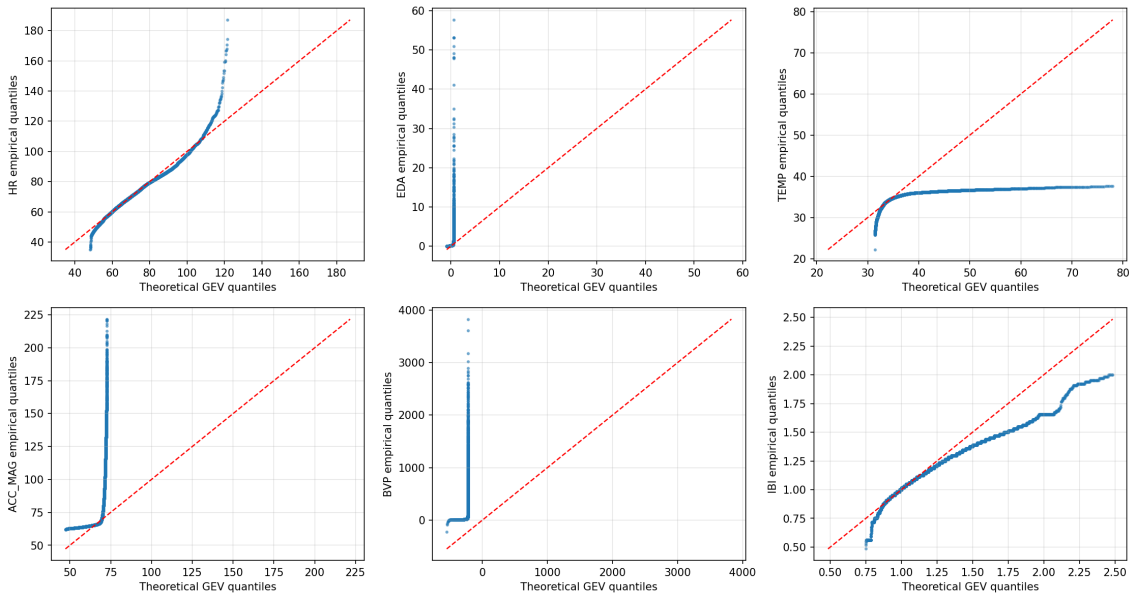

Figure 4.1 - Pooled GEV - QQ plot of fit (full population)

($3{,}493/24 \approx 145.5$ epochs) to Severe ($550/20 \approx 27.5$ epochs), this gradient is consistent with the well-documented association between OSA severity and reduced sleep efficiency (Kapur et al. 2017): participants with higher AHI tend to have shorter total sleep time and more fragmented recordings, yielding fewer scoreable 30-second epochs after artifact rejection.

Table 4.2 - GEV shape parameter $\hat{\xi}$ by signal and AHI severity class (Normal $n = 3\,493$; Mild $n = 2\,190$; Moderate $n = 1\,042$; Severe $n = 550$ epoch maxima).

Signal	Normal	Mild	Moderate	Severe
HR	0.003	0.086	−0.013	−0.137
IBI	0.246	0.008	0.125	0.234
TEMP	0.577	0.500	0.655	0.404
EDA	−0.931	−0.883	−1.181	−0.869
ACC_{mag}	−0.936	−1.065	−1.008	−1.369

The shape estimates show no consistent monotonic trend across severity classes for any signal. For HR, the shape decreases from Mild ($\hat{\xi} = 0.086$) through Moderate (-0.013) to Severe (-0.137), transitioning from the Fréchet to the Weibull domain; however, this pattern is not replicated in the other signals. TEMP remains robustly in the Fréchet domain ($\hat{\xi} \in [0.40, 0.65]$) across all severity classes, reflecting the slow thermal dynamics of wrist temperature that are unaffected by the frequency of respiratory events.

For each signal, a Kruskal-Wallis test was conducted to assess whether the median individual-participant $\hat{\xi}$ differs across the four AHI severity classes (H_0: the four severity classes share the

same population median $\hat{\xi}$, against H_1: at least one class differs). None of the six tests reached statistical significance at the 5% level (HR: $H = 3.24$, $p = 0.356$; EDA: $H = 1.83$, $p = 0.609$; TEMP: $H = 2.85$, $p = 0.415$; ACC: $H = 1.58$, $p = 0.664$; BVP: $H = 2.67$, $p = 0.445$; IBI: $H = 5.01$, $p = 0.171$). This absence of significance may reflect either a true absence of effect or limited statistical power given the sample size and observed variability, and should not be interpreted as evidence against a physiological relationship.

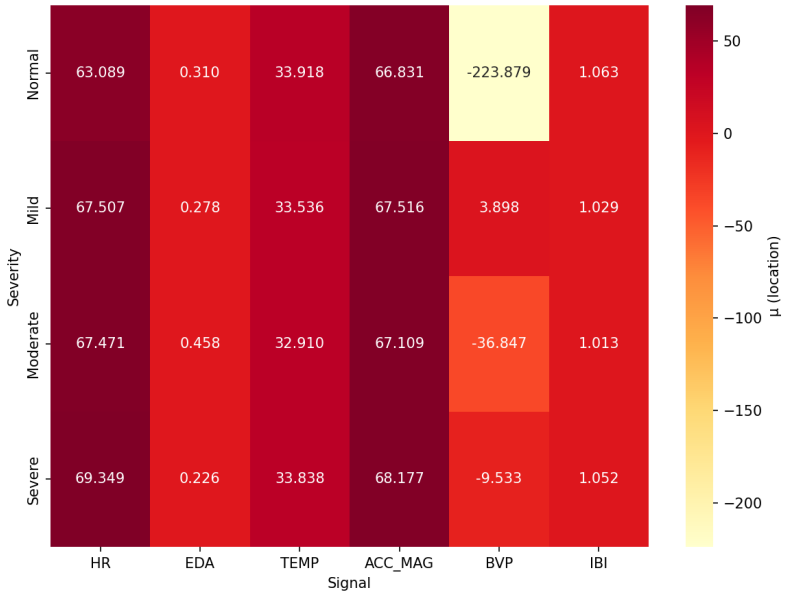

Figure 4.2 - GEV by apnea severity - parameter comparison

Figure 4.2 visualizes the full signal $\times$ severity matrix of shape estimates. Table 4.2 reports the exact numerical estimates, while Figure 4.2 is retained alongside it to highlight the visual pattern across signals and severity classes at a glance. The thermal heatmap suggests that signal membership in the Fréchet or Weibull domain is a property of the physiological signal type rather than of the participant's OSA severity.

For the three Fréchet-domain signals, the POT complement of the GEV analysis, using automatically selected thresholds at the 80th-percentile of the pooled epoch maxima, yields consistent results: HR $\hat{\xi} = 0.141$ and IBI $\hat{\xi} = 0.153$ remain in the Fréchet domain. Figure 4.3 presents violin plots of the individual participant-level POT shape estimates by AHI severity group; consistent with the GEV analysis, no group shows a dominant positive or negative shape across all signals.

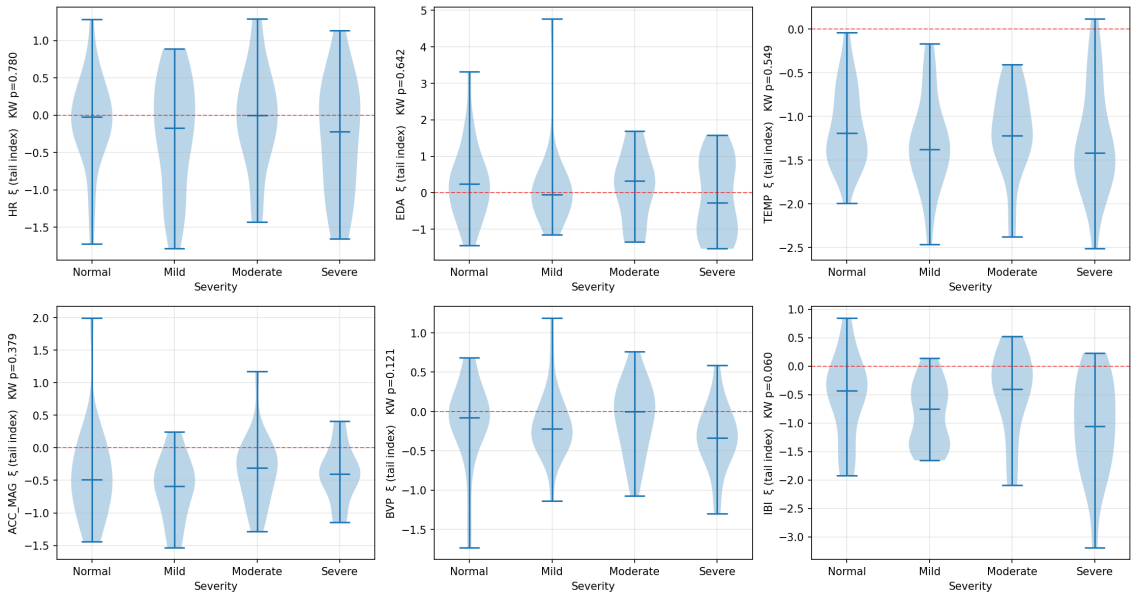

Figure 4.3 - POT shape parameter ξ by apnea severity

4.2. Population-level GPD fit

The left-truncated GPD is first fitted to the pooled declustered sample of 2 262 durations, treating all participants as exchangeable. Prior to fitting, durations are standardized by the sample median $\hat{m}$ and interquartile range $\widehat{\text{IQR}}$ of the declustered pool:

$$\tilde{d}_k = \frac{d_k - \hat{m}}{\widehat{\text{IQR}}}. \tag{4.1}$$

A data-driven exceedance threshold is selected at the 90th empirical percentile of the standardized values ($\tilde{\tau} = 1.496$). The MLE fit to the 227 exceedances yields:

$$\hat{\xi} = -0.009, \qquad \hat{\sigma} = 1.078 \quad \text{(standardized scale)}. \tag{4.2}$$

The shape estimate $\hat{\xi} = -0.009$ is indistinguishable from zero and places the pooled duration distribution near the boundary between the Gumbel and Weibull domains of attraction, rather than confirming a heavy (Fréchet) tail. Re-estimating $\hat{\xi}$ across the threshold percentile range $[0.80,\ 0.95]$ yields values ranging from $+0.043$ (80th percentile) to -0.016 (95th percentile), the sign itself is not stable within this range. This instability indicates that, at the population level, the left-truncated GPD does not provide strong evidence for a robustly heavy-tailed duration distribution: the qualitative heavy-tail signal visible in the lower-threshold diagnostics of Figure 3.1 attenuates as the threshold is raised, consistent with a tail decay closer to exponential than to a power law. This threshold-sensitivity, rather than a single point estimate, is the more accurate

population-level characterization of the duration distribution and motivates the individual-level analysis of Section 3.3, where severity is assessed participant-by-participant rather than assuming a single shared tail index.

4.3. Truncated GPD applied to apnea event durations

4.3.1. Threshold selection and justification

The left-truncation threshold $\tau = 10$ s is fixed at the AASM minimum scoring duration (cf. Section 3.2.1). To verify that this physiological threshold is also statistically appropriate, the mean excess function was evaluated over the pooled declustered sample ($n = 2\,262$ events).

As previously shown in Figure 3.2 (Section 3.2.2), the mean excess function evaluated on the same pooled declustered sample is positive and approximately linear above $u = 10$ s, consistent with the use of a GPD approximation above the clinical threshold, although the sign and magnitude of the shape parameter are threshold-sensitive (Section 4.2, where $\hat{\xi} \approx -0.009$).

4.3.2. Individual-level estimation of ξ and σ_τ

The truncated GPD was fitted separately for each of the 71 participants with at least ten declustered events ($N_i \geq 10$), using the fixed threshold $\tau = 10$ s and the MLE procedure of Section 3.2.2.

The individual shape estimates $\hat{\xi}_i$ span a wide range: approximately 14 % of participants ($n = 10$) yield $\hat{\xi}_i > 0$ (Fréchet domain), suggesting heavier-tail behavior under the fitted model than a finite physiological ceiling on event duration would predict; the remaining 86% have $\hat{\xi}_i < 0$ (Weibull domain), consistent with a finite physiological ceiling on event duration. A subset of participants hit the lower optimizer bound at $\hat{\xi}_i = -0.49$, reflecting either very few events or a strongly bounded duration distribution.

Figure 4.4 displays the distribution of individual estimates $\hat{\xi}_i$, $\hat{\sigma}_{\tau,i}$, and the model-predicted 95th-percentile $P_{95,i}$, stratified by AHI severity class.

Figure 4.5 shows the fitted truncated GPD survival curves for four representative participants, illustrating the heterogeneity in tail behavior across the cohort. Under the fitted GPD model, participants with $\hat{\xi}_i > 0$ (S037, $\hat{\xi} = 0.52$; S067, $\hat{\xi} = 0.20$) display a slowly decaying tail compatible with power-law behavior, where the fitted curve decays gradually and never reaches zero within

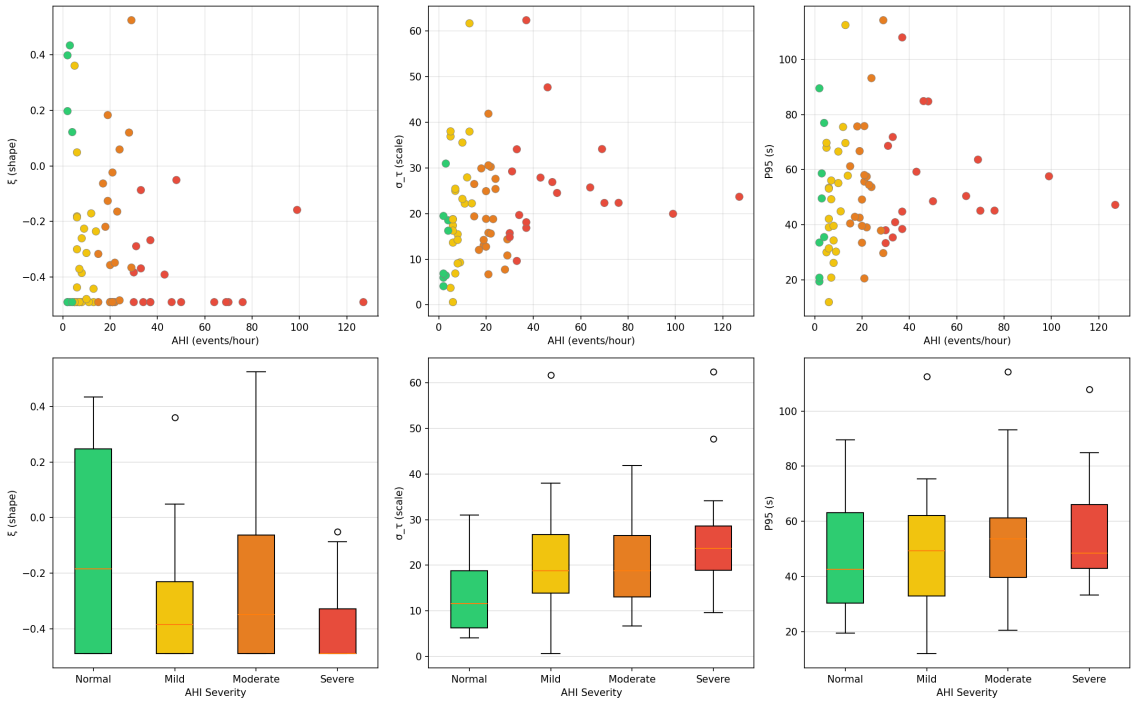

Figure 4.4 - Truncated GPD - individual fits as AHI model

the observable range; participants with $\hat{\xi}_i < 0$ (S043, $\hat{\xi} = -0.44$; S031, $\hat{\xi} = -0.39$) have a finite upper bound, visible as the steep drop and abrupt termination of their fitted survival curve.

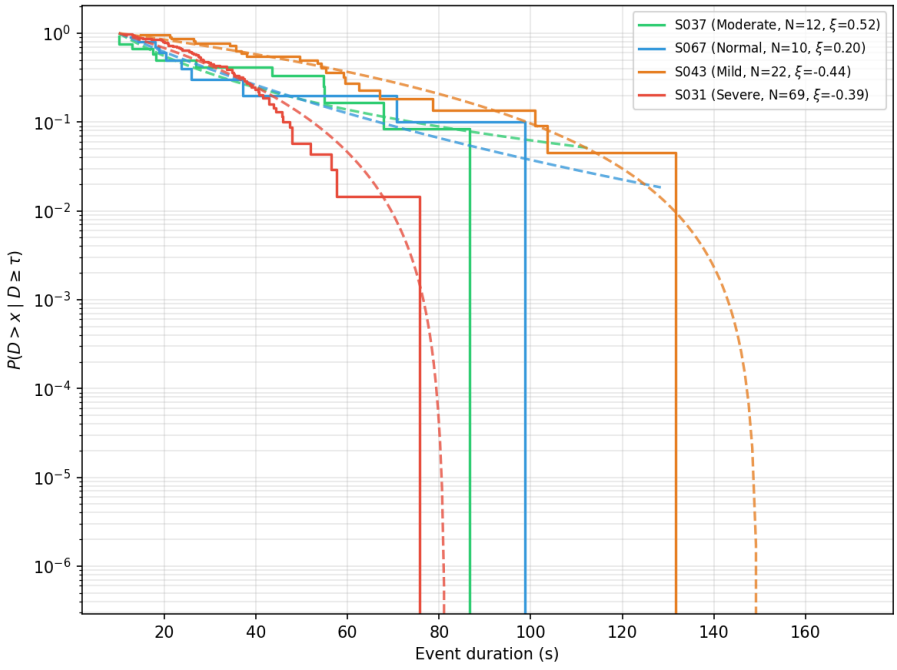

Figure 4.5 - Individual truncated GPD survival curves: empirical (solid) vs fitted (dashed)

4.3.3. Estimated extreme quantiles

The truncated GPD provides closed-form estimates of extreme quantiles of event duration for each participant via the inverse of Equation (1.2):

$$\hat{Q}_i(p) = \begin{cases} \tau + \dfrac{\hat{\sigma}_{\tau,i}}{\hat{\xi}_i}\left[(1-p)^{-\hat{\xi}_i} - 1\right], & \hat{\xi}_i \neq 0, \\ \tau - \hat{\sigma}_{\tau,i}\log(1-p), & \hat{\xi}_i = 0. \end{cases} \tag{4.3}$$

Table 4.3 summarizes the median estimated P90, P95, and P99 quantiles by AHI severity class.

Table 4.3 - Median (IQR) of truncated GPD estimated extreme quantiles by AHI severity class.

Severity	$\boldsymbol{P_{90}}$ **(s)**	$\boldsymbol{P_{95}}$ **(s)**	$\boldsymbol{P_{99}}$ **(s)**
Normal (< 5)	30 [19–57]	42.5 [20–75]	44 [22–115]
Mild (5–15)	37 [30–46]	49.2 [35–57]	52 [44–70]
Moderate (15–30)	45 [36–57]	53.7 [40–68]	68 [46–95]
Severe (≥ 30)	44 [37–62]	48.5 [43–72]	56 [51–87]

Several observations are notable. First, the Moderate severity class has the highest median P_{95} (54 s) and P_{99} (68 s), whereas the Severe class median P_{95} (49 s) falls below Moderate, echoing the non-monotonic pattern seen in the maximum duration analysis (Section 4.5.2). This pattern may reflect a combination of subgroup heterogeneity, declustering effects, possible ceiling effects, and limited sample size; therefore, it should be interpreted as exploratory rather than as evidence that Severe OSA has shorter extreme durations. Second, the Normal group exhibits very high inter-participant variability in P_{99} (IQR: 22-115 s), this group is small ($n = 8$) and includes four participants with a positive estimated shape parameter ($\hat{\xi}_i > 0$, Fréchet domain); because a positive ξ implies an unbounded, slowly decaying tail, the truncated- GPD estimate of P_{99} for these four participants is disproportionately large, which inflates the upper tail of the Normal group's P_{99} distribution. Given the small size of this group, these results should be regarded as exploratory and are likely sensitive to the inclusion or exclusion of a small number of individuals with a positive estimated shape parameter. Third, participants with $\hat{\xi}_i > 0$ can produce P_{99} estimates exceeding 150 s (e.g., S037: 288 s; S067: 156 s), values that are extrapolations beyond the observed data and should be interpreted with caution.

4.3.4. Goodness-of-fit and model diagnostics

Figure 4.6 presents the goodness-of-fit diagnostics for the population-level truncated GPD fit using the fixed threshold $\tau = 10$ s ($n = 2\,262$ declustered events), which yields $\hat{\xi} = -0.079$, $\hat{\sigma} = 14.7$ s, a second, independent population-level estimate that, like the data-driven-threshold estimate of Section 4.2 ($\hat{\xi} \approx -0.009$), is small and negative, reinforcing the conclusion that, at the population level, the distribution does not look as heavy-tailed. These two estimates are not directly comparable because they are obtained with different thresholds and different subsets of the data, but their qualitative convergence, both small and negative under independent configurations, strengthens the overall conclusion.

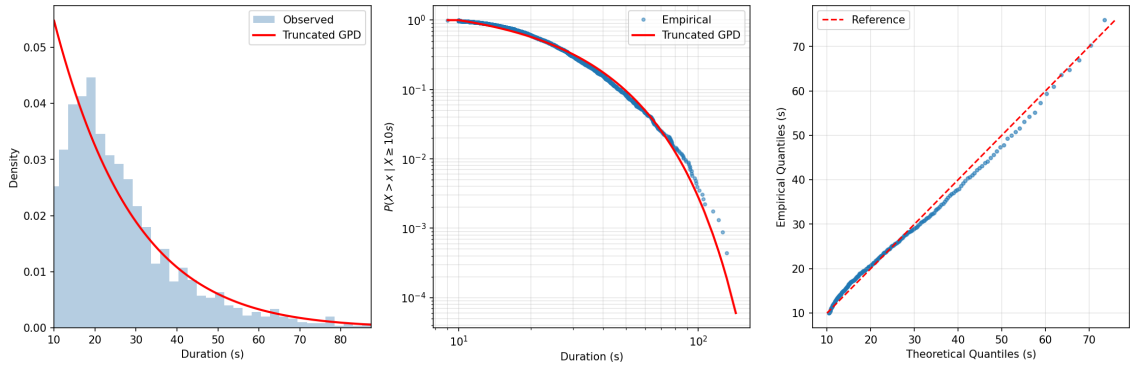

Figure 4.6 - Population-level truncated GPD goodness-of-fit diagnostics. **Left:** density overlay. **Center:** log-log CCDF **Right:** Q-Q plot

The CCDF log-log plot is visually compatible with approximately linear behavior over part of the upper tail, although the population-level shape estimate remains close to zero and threshold-sensitive (Section 4.2).

4.4. EVT parameters as indicators of apnea severity

4.4.1. Association between AHI and tail index ξ

Figure 4.7 presents the correlation matrix between individual extreme-value metrics, including the shape parameter $\hat{\xi}_i$, maximum duration, and duration percentiles, and clinical variables (AHI, OAHI, BMI, mean SpO2). OAHI restricts the AHI count to obstructive events only, excluding central and mixed apneas, and is included here as a specificity check on the AHI-based associations.

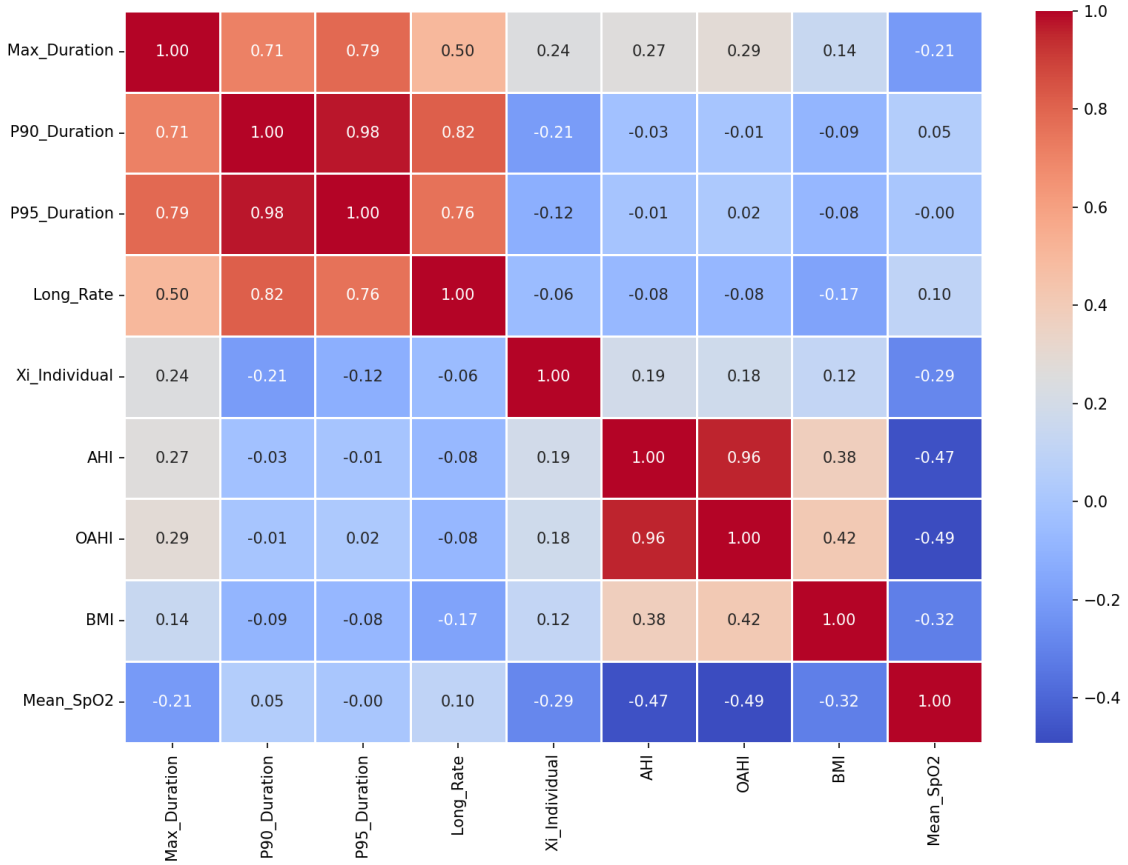

Figure 4.7 - Correlation matrix: extreme-value metrics vs clinical variables (AHI, OAHI, BMI, SpO2)

The shape parameter $\hat{\xi}_i$ shows a weak, statistically non-significant association with AHI: the Fréchet-domain participants ($\hat{\xi}_i > 0$) are distributed across all four severity classes, confirming that having a heavier-tailed duration distribution is not exclusively a property of severe OSA. This is consistent with the Kruskal-Wallis non-significant result for the wearable signal GEV analysis (Section 4.1.2): tail heaviness, as characterized by ξ, is not a reliable marker of AHI severity in this cohort.

4.4.2. Association between AHI and scale parameter σ_τ

Table 4.4 reports the median $D_i^{\max}$ and $P_{95,i}$ stratified by AHI severity class, for the 71 participants with individual GPD fits.

The effective scale $\hat{\sigma}_{\tau,i}$ shows a stronger, positive monotonic association with AHI than $\hat{\xi}_i$ alone. Scale governs the typical magnitude of events above τ and is therefore more directly linked to the characteristic duration of respiratory events, which tends to increase with disease severity.

Table 4.4 - Median (IQR) of duration statistics by AHI severity class, for participants with at least ten declustered events.

Severity	n	$D^{\max}$ (s)	P_{95} (s)	$D^{\max}_{\text{range}}$ (s)
Normal (< 5)	8	45.9 [33.4–63.6]	42.5	19.4–98.8
Mild (5–15)	23	52.9 [36.8–70.9]	49.2	22.5–131.6
Moderate (15–30)	21	65.4 [50.3–80.8]	53.7	21.4–143.4
Severe (≥ 30)	19	50.9 [45.8–87.3]	48.5	36.8–126.1

Figure 4.8 presents the distribution of three duration severity metrics: $D_i^{\max}$, $P_{95,i}$, and the proportion of events exceeding 30 s, stratified by AHI severity class.

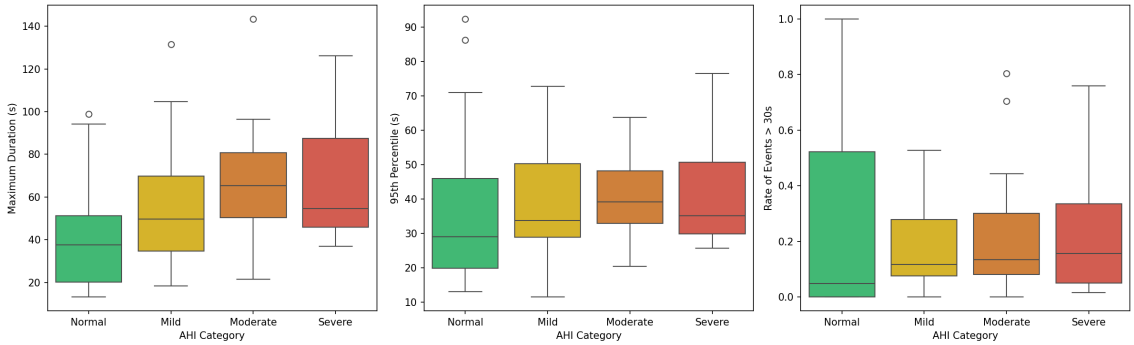

Figure 4.8 - Extreme-value metrics by AHI severity

The Spearman correlation between $D_{\max,i}$ and AHI is $\rho = 0.23$ ($p = 0.052$) and between $P_{95,i}$ and AHI is $\rho = 0.19$ ($p = 0.104$); neither reaches conventional significance. This weak association indicates that the duration tail carries information not captured by AHI frequency alone: a duration metric strongly correlated with AHI would add little beyond what AHI already captures, whereas a weakly correlated one may reflect a complementary dimension of severity. However, because these correlations are weak and do not reach conventional statistical significance, this interpretation should be regarded as hypothesis-generating; an alternative explanation is measurement noise or limited statistical power at the participant level.

A notable exception is participant S043 (AHI=13, Mild class), whose effective scale $\hat{\sigma}_\tau = 61.7$ s is among the three largest in the cohort, with a model-predicted $P_{95} = 112$ s. This participant illustrates a pattern in which individual respiratory events are extraordinarily prolonged despite a moderate event frequency, the archetypal case that the combined classifier is designed to flag.

4.5. Duration-based classification and combined risk assessment

4.5.1. Apnea duration classifier: light, normal and severe events

The duration-based score $B_i \in \{0, 1, 2\}$ assigns each participant to one of three levels according to their maximum observed event duration $D_i^{\max}$ and a threshold pair (T_1, T_2), as defined in Equation (3.6). Figure 4.9 presents the population-level distribution of all scored event durations: the empirical distribution departs markedly from both the Normal and Exponential reference densities (center panel), and is visually compatible with a fitted GPD above the AASM threshold (right panel), consistent with the right-skewed shape discussed in Chapter 3, although the population-level shape estimate itself remains weak and threshold-sensitive (Section 4.2).

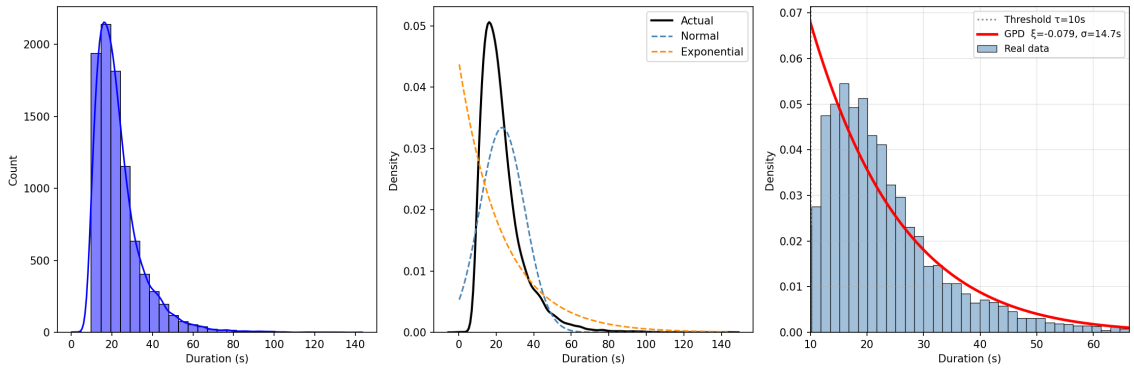

Figure 4.9 - Distribution of apnea event durations in the DREAMT cohort. **Left:** Actual duration distribution **Center:** Comparison with normal and exponential **Right:** PDF: fitted GPD vs real data

Of the 89 participants with at least one recorded apnea event, 18 (20 %) have $D_i^{\max} \leq 34$ s (score $B_i = 0$, "light"), 28 (31 %) have $34 < D_i^{\max} \leq 52$ s ($B_i = 1$, "normal"), and 43 (48 %) have $D_i^{\max} > 52$ s ($B_i = 2$, "severe"). The high proportion of participants with at least one event exceeding 52 s underscores that prolonged individual events are common even in the Normal and Mild AHI classes.

4.5.2. The two-dimensional risk space: AHI and maximum event duration

Figure 4.10 projects all 90 participants into the AHI-$D_i^{\max}$ plane.

The two-dimensional view reveals a substantial stratum of participants whose AHI-alone class underestimates their physiological risk: 18 individuals simultaneously satisfy AHI < 15 events/h (Normal or Mild class) and $D_i^{\max} > 52$ s, placing them in the upper-left region of Figure 4.10,

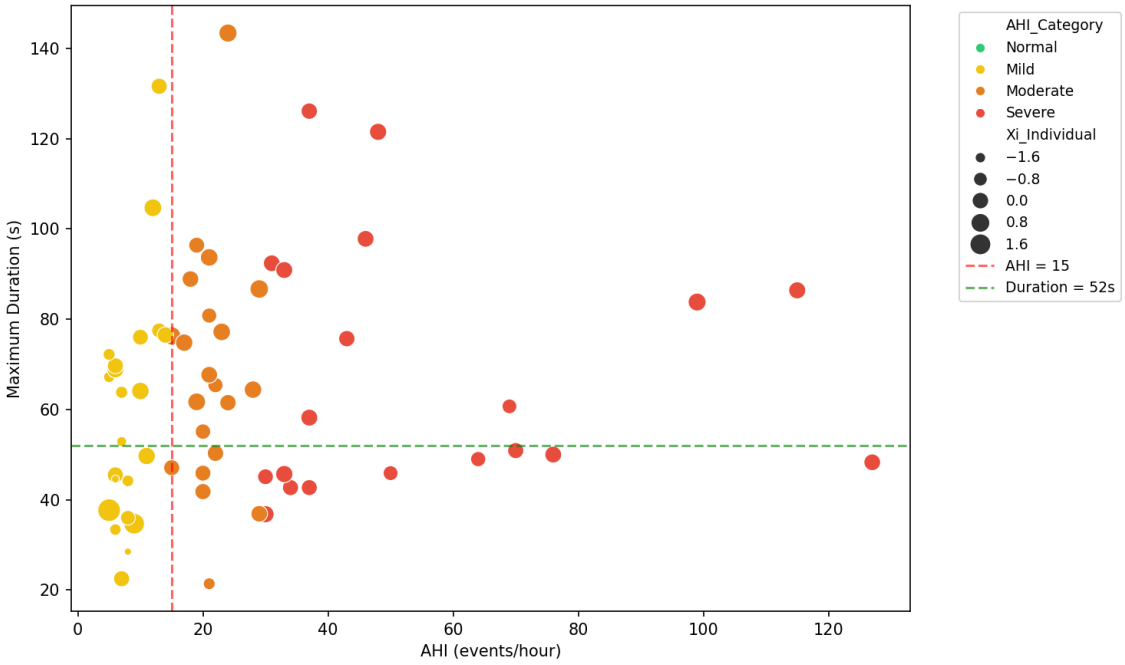

Figure 4.10 - AHI vs maximum duration (upper-left zone = undetected risk)

bounded by the two dashed reference lines. All 18 have at least one recorded apnea event long enough to be associated, in the literature, with SpO2 desaturation to $\leq 80\,\%$ (Zhou et al. 2026), yet their AHI-alone classification would suggest only mild or no pathology.

4.5.3. Combined severity classifier: design and reclassification analysis

Under the reference thresholds $(T_1, T_2) = (34\,\text{s}, 52\,\text{s})$, the combined classifier $C_i = \max(A_i, B_i)$ upgrades 24 of the 90 participants (26.7 %) to a higher severity class. No participant is downgraded (by construction, $C_i \geq A_i$). Table 4.5 shows the pre/post-reclassification counts for each severity class.

Table 4.5 - Reclassification matrix: Off-diagonal cells count participants upgraded by the duration score; the diagonal counts participants whose classification is unchanged.

	Combined class C_i				
AHI class A_i	Normal	Mild	Moderate	Severe	**Total**
Normal (< 5)	12	6	6	0	24
Mild (5–15)	0	13	12	0	25
Moderate (15–30)	0	0	21	0	21
Severe (≥ 30)	0	0	0	20	20
Total C_i	12	19	39	20	90

The most common upgrade path is Mild → Moderate ($n = 12$, representing 48 % of all Mild participants), followed by Normal → Mild ($n = 6$, 25 % of Normal participants) and Normal → Moderate ($n = 6$, 25 % of Normal). After reclassification, the Moderate class nearly doubles in size (from $n = 21$ to $n = 39$), reflecting the substantial share of participants in the Normal and Mild AHI groups who exhibit individually severe event durations.

Figure 4.11 illustrates the reclassification in terms of the compound Poisson- GPD risk quadrants defined by event rate λ_i and tail severity $\hat{\xi}_i$.

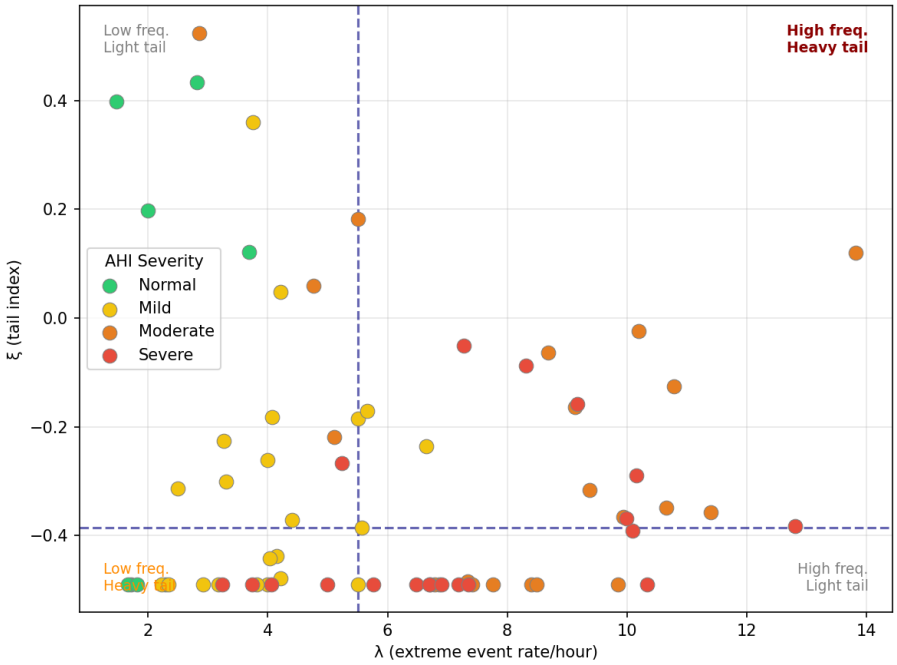

Figure 4.11 - Risk quadrants: frequency x tail index

4.5.4. Identification of undetected risk

The compound Poisson- GPD model provides a principled quantification of the duration-based risk profile carried by the undetected stratum. Figure 4.12 shows the 8-hour return level curves for the four AHI severity groups alongside the risk quadrant scatter.

Several findings emerge from Figure 4.12. First, the Moderate and Severe groups achieve the highest 8-hour return levels on average, driven primarily by their higher event rates λ_i. Second, a subset of Normal and Mild AHI participants (the low-frequency/heavy-tail quadrant, cf. Figure 4.11) exhibits 8-hour return levels comparable to or exceeding those of many Moderate participants, despite their low AHI: events are rare but, when they occur, they are exceptionally

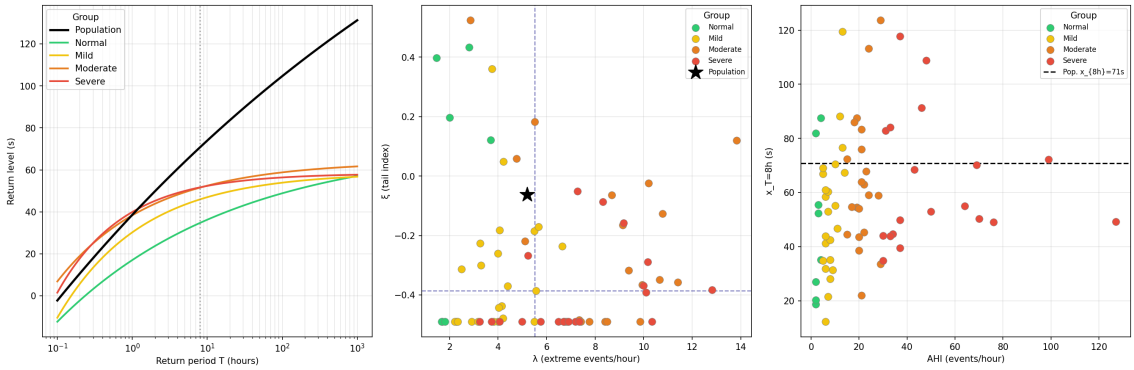

Figure 4.12 - Compound Poisson - GPD model: frequency x severity

prolonged. Third, participant S043 (AHI = 13, Mild class, $D^{\max} = 131.6$ s, $\hat{\sigma}_\tau = 61.7$ s) exhibits an 8-hour return level in the Severe range, exemplifying how a low-frequency but high-scale duration distribution (estimated scale $\hat{\sigma}_\tau = 61.7$ s, shape $\hat{\xi} = -0.442$), placing S043 in the Weibull domain with a bounded but elevated range of apnea durations.

The grid search over 11,051 threshold pairs (Table 3.1, Section 3.5.2) suggests that data-optimal thresholds modestly outperform the clinical reference pair on purely statistical grounds; the reference pair is nonetheless retained here for clinical interpretability, with the data-driven alternatives already reported as a sensitivity check in Table 3.1.

In summary, the combined classifier identifies 24 participants (26.7 %) who are upgraded relative to AHI-alone classification, of whom 18 (20 % of the cohort) fall in the potentially clinically relevant undetected-risk zone (AHI < 15, $D^{\max}$ > 52 s). The compound Poisson- GPD framework provides a theoretically grounded explanation for their elevated risk: it is not the frequency of their events that is alarming, but the tail of the duration distribution — the characteristic signature of prolonged, unresolved airway obstruction.

5. Validation and generalizability

This chapter explores the external plausibility and generalizability of the findings presented in Chapter 4. Section 5.1 compares the DREAMT cohort with two external datasets on AHI distribution, EVT tail behavior, and SpO2 patterns; since neither external dataset contains event-level duration data, Section 5.1.4 uses SpO2 as an indirect proxy to assess whether the undetected-risk phenomenon identified in Section 4.5.4 is likely to generalize beyond the DREAMT cohort. Section 5.2 presents supporting evidence from the concurrent 100 Hz PSG recordings available in DREAMT, examining epoch-level signal characterization by sleep stage, the oxygen desaturation proxy, thorax-abdomen synchrony, and consistency with the primary wearable findings.

5.1. Cross-dataset comparison

5.1.1. AHI distribution: DREAMT vs external datasets

Table 5.1 summarizes the key characteristics of the three datasets used in this project.

Table 5.1 - Comparative summary of the three sleep datasets.

Feature	DREAMT (primary)	Sleep Disorder Dataset	Sleep Health & Lifestyle
N	100 nominal / 90 usable	1 000	374
Data type	Real PSG + E4	Synthetic	Self-report
AHI available	✓	✓	×
Mean (median) age (yr)	56.6 (58.7)	40.5 ± 22.7	42.2 ± 8.7
AHI median	13.5	26.4	—
AHI IQR	[5–29]	[13.5–38.1]	—
Mean SpO_2 (%)	95	87.5	—
$SpO_2 < 90$ %	<5 %	59.9 %	—
OSA (%)	74 %	30.9 %[†]	20.9 %

[†] Obstructive Sleep apnea specifically; 100 % have some disorder.

The DREAMT severity distribution is notably more balanced than that of the Sleep Disorder Dataset (Normal 27 %, Mild 28 %, Moderate 23 %, Severe 22 % vs Normal 10 %, Mild 18 %, Moderate 31 %, Severe 41 %). This balance is consistent with a controlled sleep-laboratory study recruiting across the severity spectrum, whereas the Sleep Disorder Dataset reflects a referral-biased population with predominantly Moderate-Severe OSA.

Figure 5.1 compares the AHI distributions of DREAMT and the Sleep Disorder Dataset at the level of the empirical histogram, boxplot by severity, and CCDF.

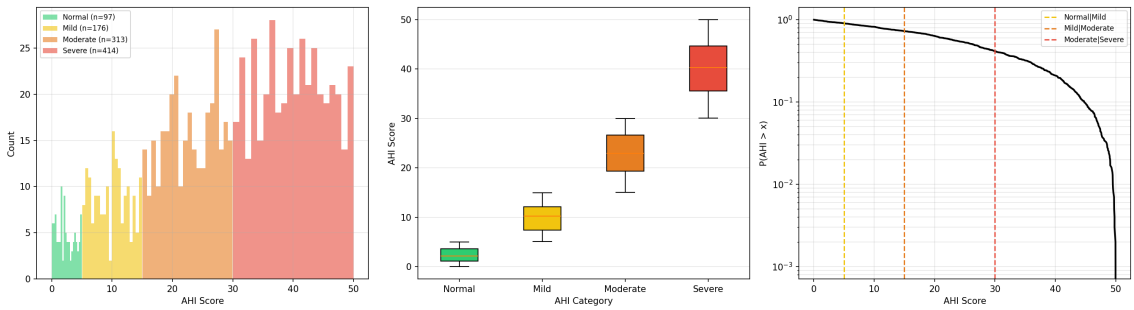

Figure 5.1 - Sleep disorder dataset - AHI score distribution.

Figure 5.2 shows the breakdown by disorder type in the Sleep Disorder Dataset: of the 1 000 records, 309 (30.9 %) are labeled Obstructive Sleep apnea, 255 (25.5 %) Insomnia, 203 (20.3 %) Restless Leg Syndrome, 150 (15.0 %) Narcolepsy, and 83 (8.3 %) have no recorded disorder. Importantly, AHI scores are present for all disorder types (not only OSA), and the administrative provenance of the dataset (uniform age distribution 3-80 yr, templated diagnostic text) means that cross-dataset AHI comparisons must be treated as indicative rather than clinically definitive.

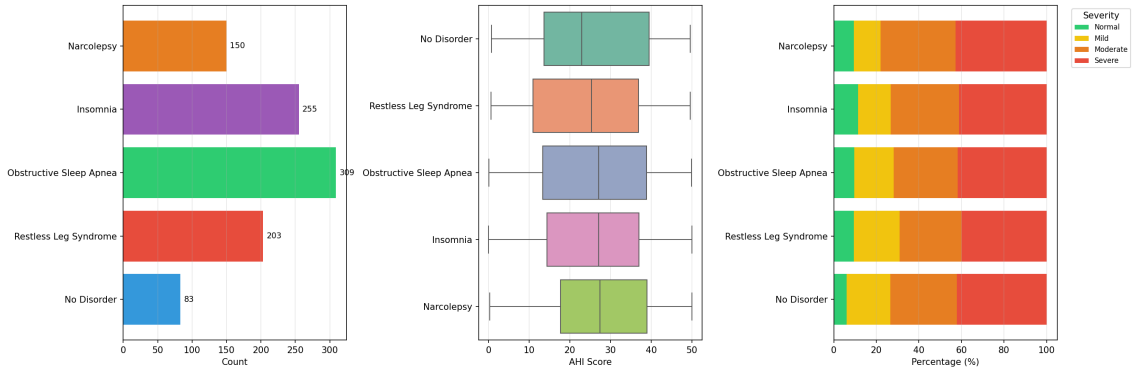

Figure 5.2 - Sleep Disorder Dataset - distribution by disorder type and AHI.

5.1.2. EVT applied to AHI score across populations

The CCDF log-log plot in Figure 5.1 (right panel) reveals whether the tail of the AHI distribution is heavy across populations. In the Sleep Disorder Dataset, the empirical CCDF above AHI = 30 approximates a linear decay on the log-log scale, suggesting that the right tail of the AHI distribution is consistent with a power-law, placing it in the Fréchet domain of attraction. This is analogous to the behavior observed for apnea event durations in DREAMT (Section 4.2).

The DREAMT AHI tail is substantially lighter: only 22 % of participants exceed AHI = 30, whereas in the Sleep Disorder Dataset 41 % are in the Severe class (AHI ≥ 30, maximum = 49.99 events/h). This difference reflects the sampling design rather than a fundamental difference in tail behavior: a POT applied to the Sleep Disorder Dataset AHI scores above threshold $u = 15$ events/h could be assessed to formally estimate the shape parameter, which the right-skewed, fat-tailed shape of the distribution suggests would be positive, consistent with OSA severity in referral populations.

The key implication for the EVT framework developed in this project is that duration-based tail behavior can be assessed independently of the AHI: the AHI score itself exhibits right-tail behavior consistent with the Fréchet domain across independent populations, but this population-level result for AHI is not evidence that individual apnea event durations follow the same tail behavior, since the pooled distribution of durations in this project did not exhibit a comparably heavy tail (Section 4.2).

5.1.3. SpO_2 comparison by AHI severity category

Figure 5.3 presents the SpO_2 analysis for the Sleep Disorder Dataset.

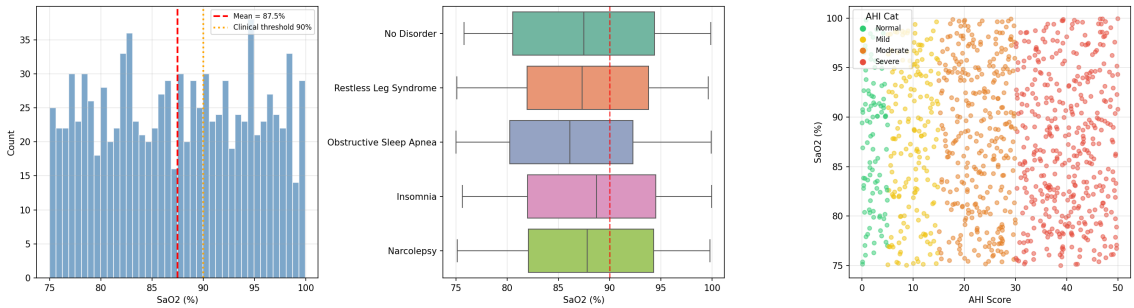

Figure 5.3 - SpO2 distribution and AHI relationship in the Sleep Disorder Dataset

The mean SpO_2 of 87.5 % in the Sleep Disorder Dataset contrasts sharply with the DREAMT cohort median of 95 % (from PSG clinical records). Three factors account for this difference:

1. **Severity composition**: the Sleep Disorder Dataset has 41 % Severe vs 25 % in DREAMT, so a greater proportion of participants have chronically depressed SpO_2.
2. **Administrative provenance**: the synthetic nature of the Sleep Disorder Dataset means the SpO_2 values may not reflect a clinically calibrated measurement, reducing direct comparability.
3. **Non- OSA disorders**: Restless Leg Syndrome and Narcolepsy records also carry low SpO_2 values, which is physiologically atypical for those conditions.

Despite these caveats, the negative monotonic association between AHI and SpO_2 is preserved across both datasets: higher AHI is consistently associated with lower mean SpO_2, supporting the construct validity of the AHI variable in the Sleep Disorder Dataset and confirming the expected physiological relationship. This association also motivates, but does not directly determine, the clinical thresholds $T_1 = 34$ s and $T_2 = 52$ s used in the combined classifier: both thresholds were derived from the duration- SpO2 relationship reported by (Zhou et al. 2026), not from the AHI- SpO2 relationship observed in the present cross-dataset comparison.

5.1.4. Transfer of the undetected risk concept

The Sleep Disorder Dataset does not contain event-level duration data, precluding a direct application of the combined frequency-severity classifier (Section 3.5). However, the SpO_2 data permit an indirect assessment of the undetected-risk hypothesis.

Among the 273 records in the Sleep Disorder Dataset with AHI < 15 (Normal: $n = 97$, Mild: $n = 176$), a substantial proportion have SpO_2 values below 90 %: overall, 59.9 % of all records fall below this threshold, suggesting that a non-trivial fraction of Normal/Mild-classified patients experience clinically significant hypoxemia despite their low event frequency. If the duration-based combined classifier were applicable, some of these patients would be expected to fall in the undetected-risk stratum analogous to the 18 DREAMT participants identified in Section 4.5.4.

The Sleep Health & Lifestyle Dataset (Figure 5.4) provides a lifestyle-level perspective: among the 78 self-reported Sleep Apnea cases (21 %), the dataset reports higher blood pressure, higher BMI, and lower sleep quality relative to the None group, consistent with the known comorbidity profile of OSA. The absence of AHI or event-level data in this self-report dataset limits formal

EVT analysis; the Sleep Apnea group's elevated cardiovascular risk markers support the general plausibility of comorbidities associated with OSA, but do not validate the specific hypothesis of prolonged individual events or undetected risk

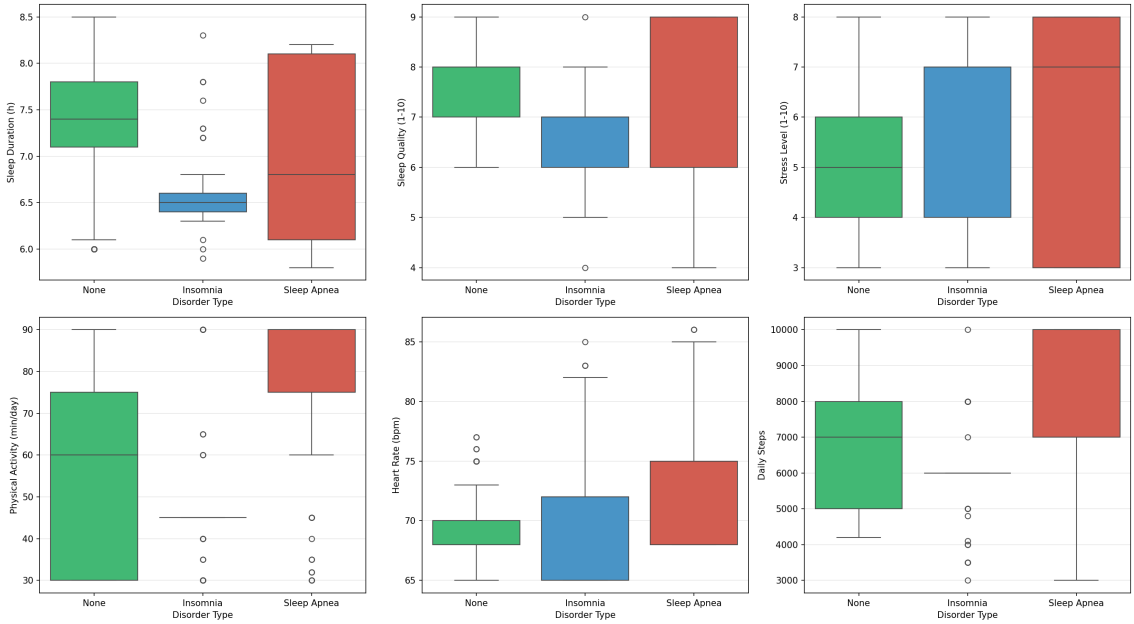

Figure 5.4 - Sleep Health & Lifestyle dataset - sleep metrics by disorder.

5.2. Supporting evidence from 100 Hz PSG signals

5.2.1. Epoch-level signal characterization by sleep stage

The concurrent 100 Hz PSG recordings available in DREAMT comprise five channels: nasal pressure airflow (PTAF), oronasal thermal airflow (FLOW), thoracic belt effort (THORAX), abdominal belt effort (ABDOMEN), and pulse oximetry (SaO2). The SaO2 channel is an analog transducer output in the range $[-0.14, 1.13]$, not calibrated SpO2, and cannot be directly compared with clinical desaturation thresholds without participant-level normalization (see Section 5.2.2).

The E4 wristband signals (HR, IBI, EDA, TEMP, BVP) also available at 64 Hz were characterized by sleep stage in the primary epoch-level analysis. Figure 5.5 shows the distribution of sleep stages across the cohort.

Figure 5.6 shows the grand-mean of each E4 signal by sleep stage epoch, aggregated across all participants.

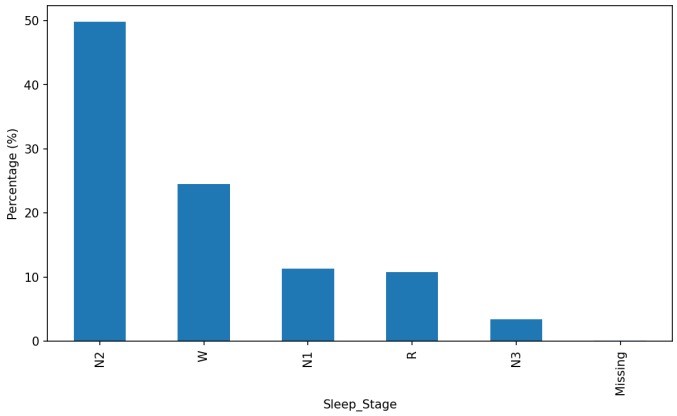

Figure 5.5 - Sleep stage distribution of scored epochs across the DREAMT cohort (percentage of total epochs).

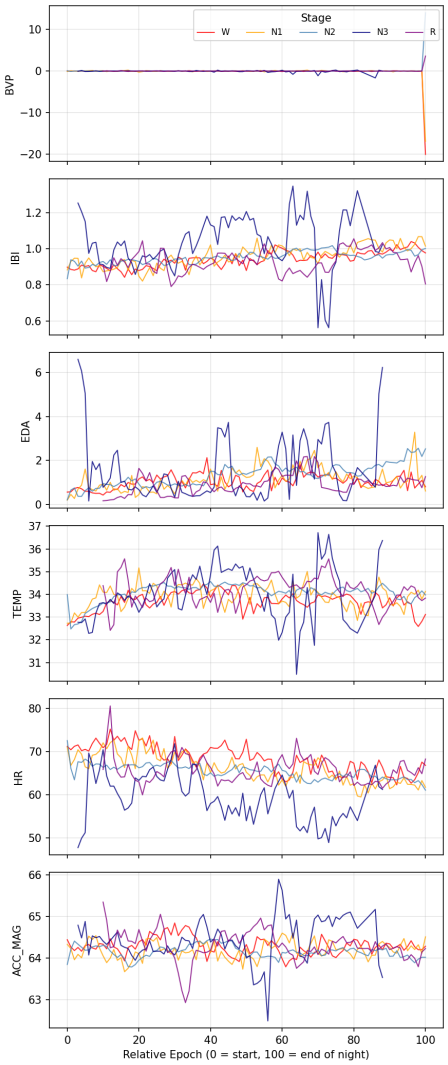

Figure 5.6 - Time series by sleep stage (global mean).

Figure 5.7 stratifies the E4 epoch-level means by AHI severity class, revealing how the wearable signal baseline differs across the clinical spectrum.

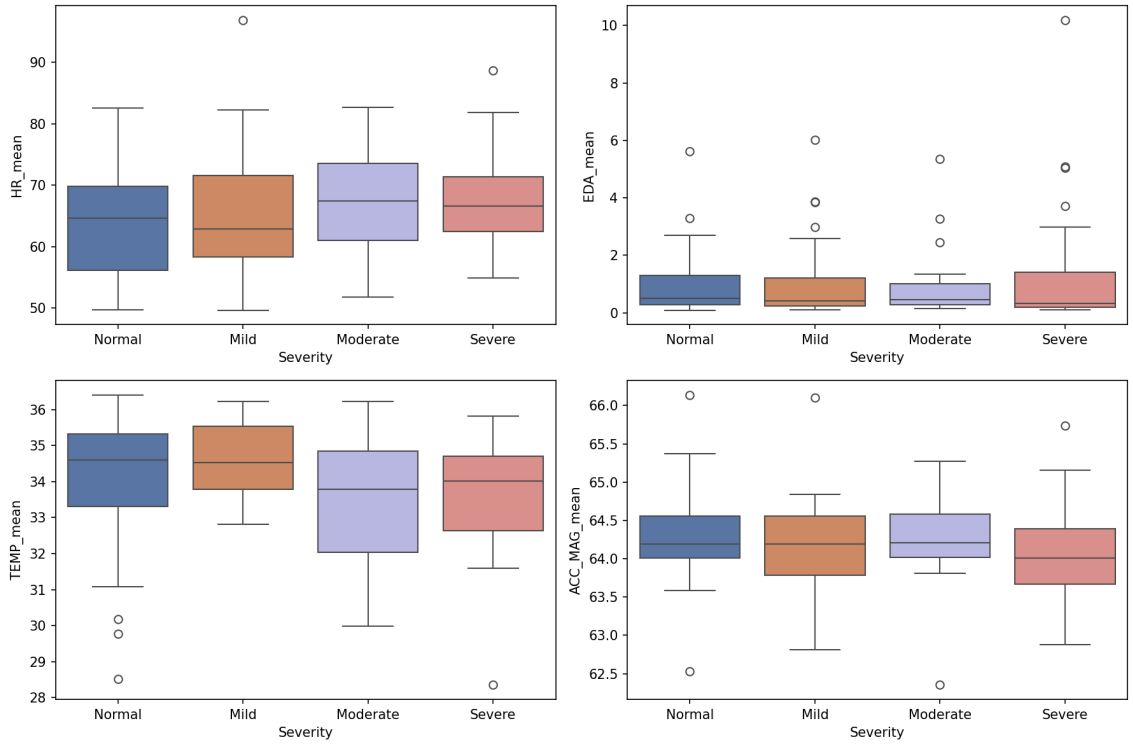

Figure 5.7 - E4 signals by apnea severity.

The E4 inter-signal Pearson correlation matrix (Figure 5.8) confirms that IBI and HR are negatively correlated ($r \approx -0.63$; by definition), while EDA and TEMP show moderate positive correlation, and ACC magnitude is weakly correlated with all other signals.

5.2.2. Temporal SAO_2 analysis and ODI3-like proxy

Because the 100 Hz SaO2 channel is an analog signal (pooled descriptive statistics: mean = 0.72, median = 0.94, range $[-0.14, 1.13]$, $n = 15\,685\,200$ samples across all participants), direct application of clinical SpO_2 thresholds (90 %, 85 %) is not meaningful without calibration. A participant-level relative normalization was therefore adopted:

$$\widetilde{s}(t) = \frac{s(t) - s_{\min}}{s_{\max} - s_{\min}}, \qquad \widetilde{s}(t) \in [0, 1], \tag{5.1}$$

where $s_{\min}$ and $s_{\max}$ are the participant's minimum and maximum raw SaO2 values. A baseline oxygenation level is estimated as the 95th percentile of $\tilde{s}(t)$ across the recording ($\hat{b}_{95}$), chosen as a level high enough to represent resting (non-desaturated) oxygenation while remaining robust to

Figure 5.8 - Correlation between E4 signals.

occasional sensor noise, in contrast to the maximum (P100), which would be sensitive to single-sample artifacts. An epoch of 30 s is classified as a *desaturation epoch* if its minimum normalized value falls more than 3 % below the baseline:

$$\text{desat}_e = \mathbf{1}\left\{\hat{b}_{95} - \min_{t \in e} \tilde{s}(t) > 0.03\right\}. \tag{5.2}$$

The resulting relative ODI3-like proxy is computed as the number of desaturation epochs per hour of recording:

$$\widehat{\text{ODI3}}_i = \frac{\sum_e \text{desat}_{e,i}}{T_i}, \tag{5.3}$$

where T_i is the total recording duration in hours. This proxy is analogous to the standard Oxygen Desaturation Index at the 3 % threshold (ODI3), but operates in relative rather than absolute SpO_2 units.

Figure 5.9 shows the raw SaO2 signal distribution per participant, confirming wide inter-participant variability in the analog baseline, a necessary precondition for the participant-level normalization.

Figure 5.10 shows the pooled distribution of the raw SaO2 signal.

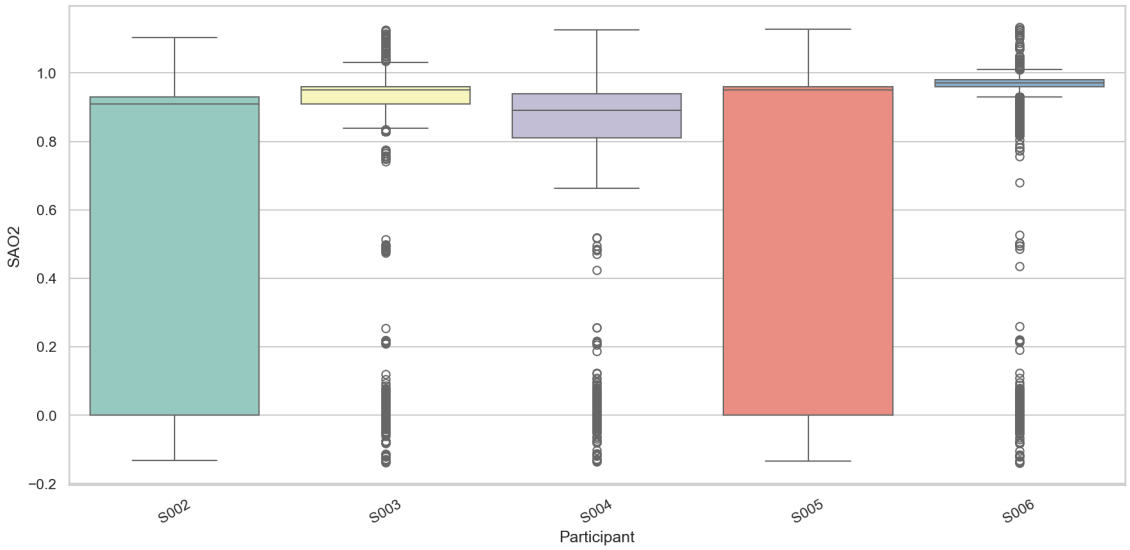

Figure 5.9 - SaO2 variation by participant (top 5).

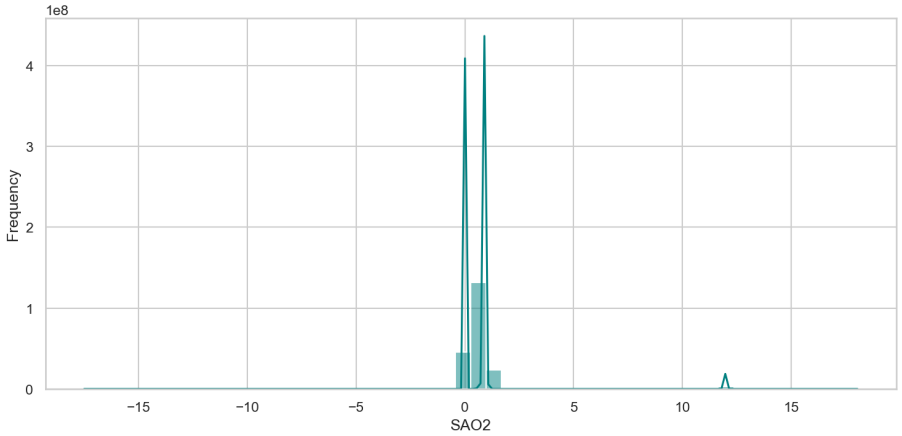

Figure 5.10 - Distribution of the raw SaO2 signal (analog transducer units, not calibrated SpO2).

5.2.3. Respiratory effort and thorax-abdomen synchrony

The THORAX and ABDOMEN belt signals provide continuous respiratory effort data at 100 Hz, enabling characterization of the respiratory drive pattern across the night. In normal breathing, both belts move synchronously (positive correlation); in obstructive respiratory events with intact central drive, the two belts move paradoxically in opposite directions, reflecting thoracic expansion attempting to overcome an occluded airway while abdominal musculature compensates in the opposite phase.

The within-participant Pearson correlation between THORAX_mean and ABDOMEN_mean per epoch was computed as a proxy for respiratory synchrony:

- $r > 0.7$: synchronized breathing (normal);
- $0.3 < r \leq 0.7$: partial synchrony (transitional);
- $r \leq 0.3$: paradoxical pattern (obstructive effort).

Figure 5.11 shows the THORAX signal distribution per participant, and Figure 5.12 presents the inter-signal correlation matrix for all five PSG channels.

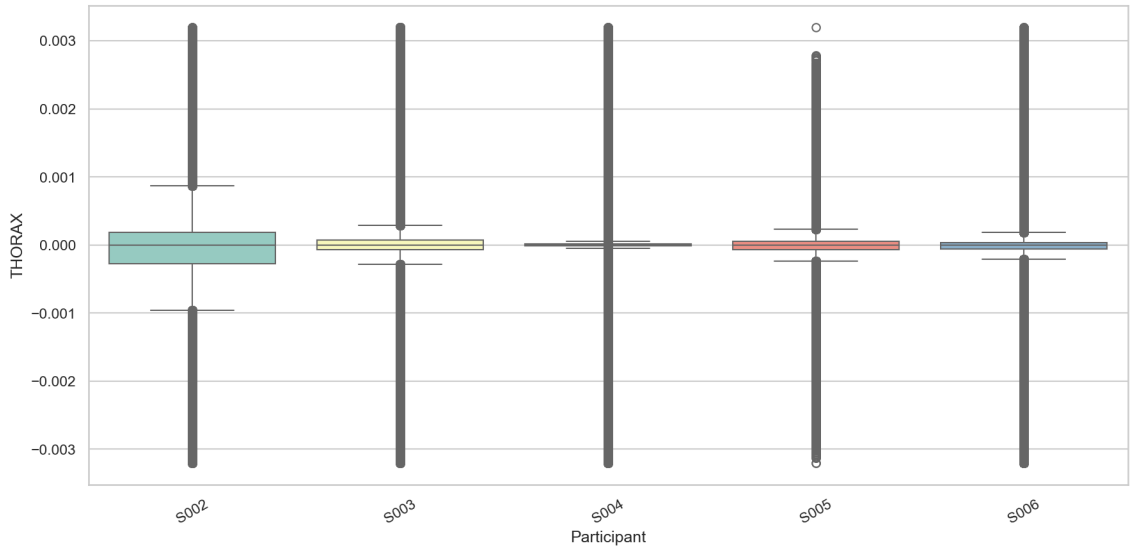

Figure 5.11 - Raw THORAX signal distribution per participant.

The pooled THORAX-ABDOMEN correlation (Figure 5.12) provides a population-level summary, but the clinically meaningful information resides at the participant and epoch level: a participant whose THORAX-ABDOMEN correlation is negative during hypnogram-confirmed NREM sleep is exhibiting paradoxical breathing characteristic of obstructive events, independent of the AHI count. This analysis is exploratory: a full quantitative synthesis (the proportion of participant-epochs exhibiting $r \leq 0.3$, stratified by AHI severity) was not performed and is left for future work. This reinforces the core project finding: the temporal pattern and duration of individual events, not only their frequency, must be characterized to fully assess OSA severity.

5.2.4. Consistency with primary dataset findings

The evidence from both the cross-dataset comparison and the 100 Hz PSG analysis is broadly consistent with the primary findings from the 64 Hz wearable analysis.

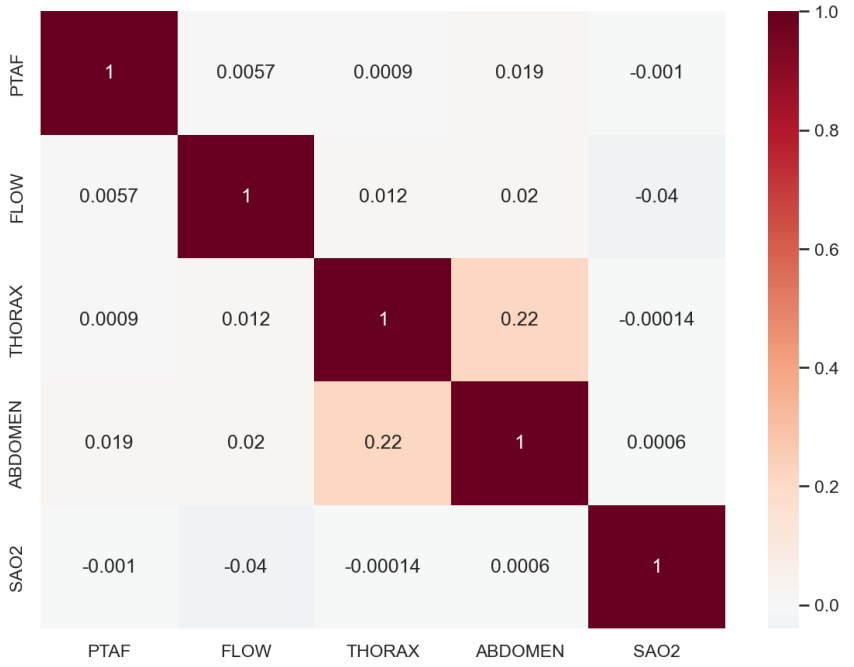

Figure 5.12 - Pearson correlation matrix between the five 100 Hz PSG channel means (PTAF, FLOW, THORAX, ABDOMEN, SaO2).

AHI-SpO_2 relationship. The negative association between AHI and mean SpO_2, confirmed in the Sleep Disorder Dataset ($\rho < 0$, $p \ll 0.001$), is the same relationship that underlies the clinical thresholds $T_1 = 34$ s and $T_2 = 52$ s used in the combined classifier. The finding that both DREAMT (PSG-derived SpO_2 from `participant_info.csv`) and the external dataset support this relationship increases confidence in the clinical grounding of the duration-based reclassification.

Tail behavior of the AHI distribution. The Fréchet-compatible tail of the AHI CCDF observed in the Sleep Disorder Dataset (Section 5.1.2) is qualitatively similar to the heavy-tail diagnostics observed for event durations in DREAMT at low thresholds (Section 3.1), although the population-level duration shape estimate itself remained weak and threshold-sensitive (Section 4.2). This suggests that both frequency and duration may exhibit extreme-event structure, although the physiological mechanisms linking these dimensions require further investigation.

Sleep stage and wearable signal profiles. The stage-stratified E4 epoch means (Figure 5.6) confirm that HR is systematically reduced during N3 and REM relative to Wake across all AHI severity groups, consistent with published normative sleep physiology (Carskadon and Dement 2011). The severity-stratified epoch means (Figure 5.7) further suggest that the Severe AHI group has elevated HR and EDA during sleep, consistent with the sympathetic hyperactivation

that accumulates from repeated subcortical arousals in untreated OSA. These findings validate the use of E4 wearable signals as physiologically meaningful proxies for sleep-state and severity-related autonomic modulation, supporting the broader project premise that wearable data contain extractable severity information beyond event frequency.

6. Final remarks and conclusion

6.1. EVT as a framework for apnea severity assessment

The central methodological contribution of this project is the application of EVT to the statistical characterization of individual apnea event durations. The Pickands–Balkema–de Haan theorem guarantees that, under regularity conditions, exceedances above a sufficiently high threshold follow a GPD (Coles 2001), and the empirical evidence presented in Chapter 4 supports the practical use of a GPD approximation for modeling upper event durations, although it does not establish a clear Fréchet-type heavy tail at the population level. The mean excess plot is approximately linear above $\tau = 10$ s, but the population-level shape estimates remain close to zero and threshold-sensitive. The population-level shape estimate $\hat{\xi} \approx -0.009$ places the pooled duration distribution near the boundary between the Gumbel and Weibull domains of attraction, rather than firmly within the Fréchet domain; re-estimation across the threshold percentile range $[0.80, 0.95]$ shows that the sign of $\hat{\xi}$ is not stable, indicating that population-level evidence for heavy-tailedness is weak and threshold-dependent (Section 4.2).

The choice of a left-truncated GPD at the AASM minimum $\tau = 10$ s (Kapur et al. 2017) is a key design decision that merits discussion. By conditioning the likelihood on $X \geq \tau$ rather than on exceedances above a data-driven threshold, the model retains all scored events, maximizes the effective sample size per participant, and produces estimates with a direct clinical interpretation: the fitted parameters describe the full distribution of *observable* respiratory events, not only the most extreme ones. The price is that the threshold is fixed rather than selected to minimize bias, so the asymptotic GPD approximation may be less accurate in the lower tail of the exceedance distribution; the goodness-of-fit diagnostics in Section 4.3.4 confirm that this has no practical consequence for the range of durations most relevant to clinical classification (> 30 s).

The extension to individual-level GPD fitting reveals substantial heterogeneity in tail behavior across participants: approximately 14 % have $\hat{\xi}_i > 0$ (Fréchet domain, unbounded tail) while

the remaining 86 % have $\hat{\xi}_i < 0$ (Weibull domain, bounded tail). This heterogeneity is clinically meaningful, it distinguishes participants whose airway obstruction tends to self-resolve within a predictable ceiling from those who occasionally sustain arbitrarily prolonged events, but does not correlate significantly with AHI severity, confirming that frequency and duration heterogeneity are partially independent dimensions of OSA severity.

The GEV analysis of the six E4 wearable signals (Section 4.1) provides a complementary perspective: epoch maxima of HR, IBI, and TEMP are in the Fréchet domain at the population level, confirming that extreme physiological responses during sleep also follow heavy-tailed distributions. The absence of a statistically significant association between the GEV shape and AHI severity (all Kruskal–Wallis $p > 0.17$; Table 4.2) suggests that the shape of the autonomic tail at the epoch level is not a reliable severity biomarker, while the *scale* and *location* parameters, which govern the magnitude rather than the tail heaviness, may carry more information and represent an avenue for future feature engineering.

The compound Poisson- GPD risk model (Section 3.4) integrates event frequency λ_i and tail severity $(\hat{\xi}_i, \hat{\sigma}_{\tau,i})$ into a single return-level metric $x_{8h,i}$ that quantifies the expected worst-case duration in a typical overnight recording. This framing is potentially clinically interpretable: a physician who sees $x_{8\text{h}} > 80$ s (Figure 4.12) could infer that the patient is likely to experience at least one event long enough to produce clinically significant hypoxemia during the night, regardless of the event count, pending independent clinical validation. The return level thus serves as a continuous severity metric that complements the ordinal AHI class by capturing event frequency and duration severity jointly.

6.2. Clinical value of the combined classifier over AHI alone

The combined frequency-severity classifier $C_i = \max(A_i, B_i)$ upgrades 24 of the 90 participants (26.7%) to a higher severity class relative to the AHI-alone classification (Table 4.5). This reclassification rate is substantial: more than one in four patients would receive a different, and more severe, classification if the duration of their longest respiratory event were taken into account alongside their event frequency.

The 18 participants in the undetected-risk stratum (AHI < 15, $D^{\max} > 52$ s; Section 4.5.2, Figure 4.10) are clinically the most significant subgroup. By the standard AHI criterion, these 18 patients would be classified as Normal or Mild and would typically not qualify for CPAP therapy under most clinical guidelines, which recommend treatment for AHI ≥ 15 or for symptomatic

patients with AHI ≥ 5 (Kapur et al. 2017). Yet each of these 18 patients has had at least one recorded event long enough to produce $SpO_2 \leq 80\,\%$ by the physiological relationship of (Zhou et al. 2026). Two cases from Table C.1 are particularly illustrative: participant S067 (AHI = 2, Normal class, $D^{\max} = 98.8\,$s) and participant S065 (AHI = 1, Normal class, $D^{\max} = 94.1\,$s) would both be classified as clinically normal under standard AHI criteria, yet each experienced at least one event exceeding 90 seconds, long enough to produce $SpO_2 \leq 80\,\%$ according to (Zhou et al. 2026). These cases exemplify the core failure mode that the combined classifier addresses: patients whose event frequency is negligible but whose individual event intensity is extreme. If such events recur across nights, they may contribute to a cumulative hypoxemic burden with potential consequences for cardiovascular and neurocognitive health (Kendzerska et al. 2014), even if the nightly event count does not trigger the clinical threshold.

The performance comparison between the reference thresholds and the data-driven optimal pair (Table 3.1) illustrates a tension inherent in translational research: data-optimal thresholds modestly improve discriminative performance, but lack the physiological grounding of the reference pair, whose thresholds are tied to specific SpO_2 nadir values. The reference thresholds are therefore retained as the primary specification, with the data-driven alternatives reported as sensitivity analyses. This choice reflects a general principle in clinical decision support: a rule that a clinician can explain to a patient is more likely to be adopted than a rule that is merely optimal on a training set. Critically, the combined classifier is not intended as a replacement for PSG, even at its best-performing threshold pair: it is a supplementary severity signal that identifies patients whose AHI-alone classification is likely to underestimate their clinical risk, triggering closer review or follow-up PSG for patients with low AHI but long maximum event durations.

6.3. Generalizability across populations and data sources

The cross-dataset analysis of Chapter 5 provides partial support for the generalizability of the findings. The key structural finding, that the AHI distribution has a heavy right tail consistent with the Fréchet domain, is replicated in the Sleep Disorder Dataset ($n = 1\,000$; Figure 5.1), suggesting that the power-law tail behavior observed for event durations in DREAMT is not a local artifact but reflects a general property of OSA severity distributions. The negative AHI-SpO2 association, which underpins the clinical justification for the duration thresholds, is also preserved across datasets.

The Sleep Health & Lifestyle Dataset ($n = 374$), although it contains no PSG data, suggests that the cardiovascular and physiological profile of self-reported Sleep Apnea is consistent with untreated OSA: elevated blood pressure, higher BMI, lower sleep quality, and elevated resting heart rate. This consistency supports the premise that the combined classifier's target population, patients with physiologically severe individual events despite low event frequency, is likely to exist in diverse real-world settings and not only in a controlled sleep-laboratory cohort.

The 100 Hz PSG analysis (Section 5.2) supports several secondary findings from the wearable epoch analysis. The E4 signal profiles by sleep stage are physiologically coherent: HR and IBI follow the expected parasympathetic dominance during NREM sleep, EDA is elevated during REM consistent with emotional arousal, and ACC magnitude discriminates wake from sleep effectively (Figure 5.6) (Carskadon and Dement 2011). The THORAX–ABDOMEN synchrony analysis (Figure 5.12) suggests that paradoxical respiratory patterns are detectable from the belt signals and are consistent with the obstructive rather than central etiology of most events in the cohort (72 % hypopneas, 15 % obstructive apneas).

Taken together, the cross-source evidence suggests that the EVT framework and the combined classifier are likely to generalize beyond the DREAMT cohort, but prospective clinical validation on an independent PSG dataset with full event annotations and clinical outcomes data remains an essential next step before the approach could be considered for clinical deployment.

6.4. Limitations

Several limitations of this work should be acknowledged.

Sample size and power. The overall reclassification analysis uses 90 participants, whereas individual GPD fitting is restricted to 71 participants with at least ten declustered events. While the full cohort is adequate for demonstrating proof of concept, it is insufficient for stable estimation of the tail index in participants with few events: of the 71 participants with $N_i \geq 10$ declustered events, a substantial fraction hit the lower bound constraint $\hat{\xi}_i = -0.49$ (Table B.1), indicating that the individual shape estimate is data-limited. A larger cohort of 300-500 participants would allow reliable individual EVT inference and formal hypothesis testing of the reclassification rate.

Single-site laboratory setting. The DREAMT dataset (Wang et al. 2024) was collected in a single sleep laboratory under controlled conditions. Laboratory PSG produces higher-quality

signal recording but may not reflect the event patterns seen in ambulatory or home-based recordings, where movement artifacts, variable sleep position, and incomplete data collection could alter the duration distribution.

No clinical outcome data. The primary validation criterion used in this project is the Spearman correlation and AUC against the continuous AHI. The AHI is itself an imperfect gold standard (Soori et al. 2022; Punjabi 2016): it does not capture symptom burden, cardiovascular risk, or response to CPAP therapy. Validation against clinical outcomes, excessive daytime sleepiness scales, cardiovascular event rates, or CPAP adherence, would substantially strengthen the clinical case for the combined classifier.

External dataset quality. The Sleep Disorder Dataset is of administrative and synthetic provenance, as evidenced by the uniform age distribution across 3-80 years, the templated diagnostic text in the `OCR_Extracted_Text` column, and the implausible combination of non- OSA diagnoses with very low SpO_2 values. Conclusions drawn from this dataset are therefore indicative rather than evidential.

SAO2 signal calibration. The 100 Hz PSG SaO2 channel is an analog transducer signal that cannot be directly compared with calibrated %SpO_2 values without participant-level normalization; the relative ODI3 proxy developed in Section 5.2.2 has not been formally validated against the clinically measured ODI in `participant_info.csv`. A mixed-effects regression of the proxy against clinical AHI and mean SpO2 would quantify this bias and, if acceptable, validate the proxy's use in datasets lacking calibrated SpO2, a natural extension for future work.

Wearable signal quality. The E4 wristband signals are subject to motion artifacts and poor skin contact during sleep, particularly for EDA and BVP (Wang et al. 2024). Missing-value rates were noted during preprocessing (Section 4.1), and the BVP-derived GEV estimates showed extreme negative shape values ($\hat{\xi} = -5.12$; Table 4.1), likely reflecting signal saturation rather than a physiological phenomenon. Analyses involving BVP and EDA should therefore be interpreted with caution.

6.5. Summary of contributions

This project makes the following original contributions:

1. **Left-truncated GPD model for apnea event durations.** Conditions the likelihood on the AASM minimum $\tau = 10$ s, yielding unbiased parameter estimates for the full observable duration distribution.

2. **Characterization of population-level tail behavior and its threshold-sensitivity.** The population-level shape estimate is weak and threshold-dependent, motivating individual-level analysis.

3. **Individual-level EVT profiling.** Per-participant GPD fits ($n = 71$) yield extreme quantiles (P_{90}, P_{95}, P_{99}) complementary to the AHI.

4. **Compound Poisson- GPD risk model.** Combines event frequency and tail severity into a single continuous return-level risk score x_{8h}.

5. **Duration-based combined severity classifier.** The upgrade rule $C_i = \max(A_i, B_i)$, benchmarked against 11,051 threshold pairs, reclassifies 24 of 90 participants and identifies 18 patients with potentially clinically relevant duration-defined undetected risk.

6. **EVT analysis of wearable physiological signals.** Suggests that HR, IBI, and TEMP epoch maxima lie in the Fréchet domain, while EDA and ACC lie in the Weibull domain.

7. **Cross-source validation framework.** Supports the AHI- SpO2 relationship and the heavy-tail structure of the AHI distribution across three independent populations.

6.6. Future work

Several extensions of this work are warranted.

Prospective clinical validation. The most pressing next step is validation of the combined classifier against clinical outcomes in an independent cohort with known CPAP adherence, cardiovascular event rates, and neurocognitive test scores. A substantially larger cohort, for example on the order of several hundred participants, would be needed to provide adequate power to test whether the 26.7% reclassification rate (Table 4.5) translates into measurable differences in clinical endpoints.

Ambulatory and home-based deployment. The DREAMT dataset was collected in a laboratory setting with concurrent PSG. Extending the analysis to ambulatory recordings, where only

the E4 wristband data are available and apnea events must be inferred from wearable signals rather than directly scored by PSG, would substantially expand the clinical reach of the framework. Recent advances in deep learning-based apnea detection from PPG and accelerometry (Wang et al. 2024) provide a natural starting point for the event extraction pipeline.

Temporal non-stationarity. The compound Poisson- GPD model assumes stationarity of the event process across the night. In practice, REM-related apneas tend to be longer and more clustered than NREM events (Rechtschaffen and Kales 1968), and the event rate λ_i may vary substantially between sleep stages. A non-stationary Poisson process with stage-dependent intensity, or a marked point process where the mark (duration) distribution varies by stage, would provide a richer model at the cost of additional parameters.

EVT features for machine learning severity prediction. The individual GPD parameters $(\hat{\xi}_i, \hat{\sigma}_{\tau,i})$, the return level $x_{8h,i}$, and the wearable signal GEV parameters are high-dimensional summaries of sleep physiology that have not been systematically evaluated as features in a supervised classification framework. Recent work has shown that lightweight tree-ensemble models can achieve strong performance on related multi-class sleep apnea severity stratification tasks using engineered physiological features (Silva, Marreiros, and Conceição 2025); incorporating EVT-derived parameters into a similar supervised framework, alongside standard sleep statistics, might substantially improve AHI severity prediction relative to any single metric.

Temperature as a candidate wearable severity biomarker. The TEMP signal exhibits the strongest and most consistent Fréchet tail across all AHI severity groups ($\hat{\xi} \in [0.40,\, 0.65]$; Table 4.2), suggesting that wrist skin temperature extremes during sleep carry physiological information that is stable across the disease spectrum. TEMP therefore appears promising for further exploration as a wearable severity biomarker, although no significant association between the GEV shape parameter and AHI severity was established in the present analysis (Section 4.1.2). The link between thermoregulatory dysfunction and OSA severity has been underexplored in the literature; a dedicated analysis of TEMP return levels as a function of AHI and SpO2 could clarify whether this signal provides incremental predictive value over HR and EDA.

6.7. Conclusion

This project asked whether Extreme Value Theory, applied to the distribution of individual apnea event durations, can provide a statistically sound and clinically interpretable characterization of OSA severity that complements the AHI. The evidence assembled across Chapters 3–5 supports a qualified affirmative answer. The empirical duration distribution departs markedly from Gaussian and exponential references, but the population-level shape estimate ($\hat{\xi} \approx -0.009$) is indistinguishable from zero and unstable across thresholds, placing the pooled distribution near the Gumbel–Weibull boundary rather than firmly within the Fréchet domain. Individual-level fitting for the 71 participants with at least ten declustered events reveals, however, substantial heterogeneity: roughly 14% of participants have $\hat{\xi}_i > 0$ (Fréchet domain), and the resulting duration-based extreme quantiles and maximum-duration score correlate only weakly with the AHI ($\rho = 0.19$–0.23, not significant), indicating that duration extremes carry information partially independent of event frequency. The GEV analysis of wearable epoch maxima similarly shows that HR, IBI, and TEMP lie in the Fréchet domain while EDA and ACC lie in the Weibull domain, with no significant association between tail shape and AHI severity.

Building on these findings, the combined classifier $C_i = \max(A_i, B_i)$, using clinically grounded thresholds ($T_1 = 34$ s, $T_2 = 52$ s), upgrades 24 of 90 participants (26.7%) to a higher severity class, of whom 18 (20% of the cohort) fall in an undetected-risk zone ($AHI < 15$, $D^{\max} > 52$ s) associated in the literature with $SpO_2 \leq 80\%$. This structural pattern, a heavy right tail in the AHI distribution and a preserved negative AHI- SpO2 association, is partially corroborated in two independent external datasets ($n = 1{,}000$ and $n = 374$), although direct transfer of the combined classifier was not possible without event-level duration annotations. Taken together, these results support the methodological value of incorporating event-duration extremes into OSA severity assessment: the AHI captures frequency but not intensity, and the combined classifier addresses this gap with a supplement that is simple, interpretable, and clinically motivated. Its adoption in practice would nonetheless require prospective validation against patient outcomes and larger, independent PSG cohorts before it could inform clinical decision-making.

References

Alimardani, M. and Moor, G. de (2021). “Automatic classification of sleep apnea type and severity using EEG signals”. In: *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies*, pp. 121–128. DOI: 10.5220/0010288301210128 (cited on page 2).

Benjafield, A. V. et al. (2019). “Estimation of the global prevalence and burden of obstructive sleep apnoea”. In: *The Lancet Respiratory Medicine* 7, pp. 687–698. DOI: 10.1016/S2213-2600(19)30198-5 (cited on pages 1, 7).

Brito, M., Freitas, A. C. M., and Freitas, J. M. (2016). “Tail prepivoting for the Hill estimator”. In: *Journal of Physics A: Mathematical and Theoretical* 49.19, p. 194004. DOI: 10.1088/1751-8113/49/19/194004 (cited on page 21).

Cao, W., Luo, J., Huang, R., and Xiao, Y. (2022). “The Association Between Sleep Breathing Impairment Index and Cardiovascular Risk in Male Patients with Obstructive Sleep Apnea”. In: *Nature and Science of Sleep* 14, pp. 53–60. DOI: 10.2147/NSS.S343661 (cited on page 2).

Carskadon, M. A. and Dement, W. C. (2011). “Normal Human Sleep: An Overview”. In: *Principles and Practice of Sleep Medicine*. Ed. by Kryger, M. H., Roth, T., and Dement, W. C. 5th. St. Louis: Elsevier Saunders, pp. 16–26 (cited on pages 51, 56).

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer (cited on pages 2, 12–18, 23, 53).

Dempsey, J. A., Veasey, S. C., Morgan, B. J., and O’Donnell, C. P. (2010). “Pathophysiology of sleep apnea”. In: *Physiological Reviews* 90, pp. 47–112. DOI: 10.1152/physrev.00043.2008 (cited on pages 1, 7, 8).

Ferro, C. A. T. and Segers, J. (2003). “Inference for clusters of extreme values”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65.2, pp. 545–556. DOI: 10.1111/1467-9868.00401 (cited on page 22).

Grimshaw, S. D. (1993). “Computing maximum likelihood estimates for the generalized Pareto distribution”. In: *Technometrics* 35.2, pp. 185–191. DOI: 10.1080/00401706.1993.10485040 (cited on page 22).

Hill, B. M. (1975). “A simple general approach to inference about the tail of a distribution”. In: *The Annals of Statistics* 3.5, pp. 1163–1174. DOI: 10.1214/aos/1176343247 (cited on page 20).

Kapur, V. K. et al. (2017). “Clinical practice guideline for diagnostic testing for adult obstructive sleep apnea”. In: *Journal of Clinical Sleep Medicine* 13.3, pp. 479–504. DOI: 10.5664/jcsm.6506 (cited on pages 1, 9, 20, 21, 28, 53, 55).

Kendzerska, T., Gershon, A. S., Hawker, G., Leung, R. S., and Tomlinson, G. (2014). “Obstructive sleep apnea and risk of cardiovascular events and all-cause mortality”. In: *PLOS Medicine* 11.2. DOI: 10.1371/journal.pmed.1001599 (cited on pages 2, 8, 10, 55).

Kulkas, A., Duce, B., Leppänen, T., Hukins, C., and Töyräs, J. (2017). “Severity of desaturation events differs between hypopnea and obstructive apnea events and is modulated by their duration”. In: *Sleep and Breathing* 21.4, pp. 829–835. DOI: 10.1007/s11325-017-1513-6 (cited on pages 2, 9, 16).

Lima, V. et al. (2026). “REMO: an integrated remote patient monitoring initiative to reduce clinical workload through workflow-oriented care pathways”. In: *Proceedings of the 39th IEEE International Symposium on Computer-Based Medical Systems (IEEE CBMS 2026)*. Accepted. IEEE (cited on page 12).

Muraja-Murro, A. et al. (2013). “The severity of individual obstruction events is related to increased mortality rate in severe obstructive sleep apnea”. In: *Journal of Sleep Research* 22.6, pp. 663–669. DOI: 10.1111/jsr.12070 (cited on pages 2, 10).

Punjabi, N. M. (2016). “COUNTERPOINT: Is the apnea-hypopnea index the best way to quantify the severity of sleep-disordered breathing? No”. In: *Chest* 149.1, pp. 16–19. DOI: 10.1378/chest.14-2261 (cited on pages 1, 10, 57).

Qin, H. et al. (2021). “The different facets of heart rate variability in obstructive sleep apnea”. In: *Frontiers in Psychiatry* 12. DOI: 10.3389/fpsyt.2021.642333 (cited on pages 1, 7).

Rechtschaffen, A. and Kales, A. (1968). *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*. Washington, DC: Public Health Service, US Government Printing Office (cited on page 59).

Rutikanga, J. U., Diop, A., and Uwilingiyimana, C. (2024). “Using extreme value theory approaches to estimate high quantiles for stroke data”. In: *Open Journal of Statistics* 14, pp. 150–162. DOI: 10.4236/ojs.2024.141007 (cited on pages 16, 18).

Seifpour, S., Niknazar, H., Mikaeili, M., and Nasrabadi, A. M. (2018). “A new automatic sleep staging system based on statistical behavior of local extrema using single channel EEG signal”. In: *Expert Systems With Applications* 104, pp. 277–293. DOI: 10.1016/j.eswa.2018.03.020 (cited on page 8).

Silva, V., Marreiros, G., and Conceição, L. (2025). “Lightweight tree ensembles with optimized features for five-class sleep apnea stratification”. In: *Progress in Artificial Intelligence*. Lecture Notes in Computer Science. 24th EPIA Conference on Artificial Intelligence (EPIA 2025), Faro, Portugal. Cham: Springer Nature Switzerland, pp. 355–367. DOI: 10.1007/978-3-032-05179-0_27 (cited on page 59).

Soori, R. et al. (2022). “Pitfalls of AHI system of severity grading in obstructive sleep apnoea”. In: *Sleep Science* 15.Special 1, pp. 285–288. DOI: 10.5935/1984-0063.20220001 (cited on pages 1, 2, 9, 10, 57).

Taran, S., Bajaj, V., Sinha, G. R., and Polat, K. (2021). “Detection of sleep apnea events using electroencephalogram signals”. In: *Applied Acoustics* 181, p. 108137. DOI: 10.1016/j.apacoust.2021.108137 (cited on page 8).

Thomas, M. et al. (2016). “Applications of extreme value theory in public health”. In: *PLOS ONE* 11.7, e0159312. DOI: 10.1371/journal.pone.0159312 (cited on page 2).

Vieira, P. et al. (2026). “An integrated wearable and AI ecosystem for remote elderly monitoring and medication adherence”. In: *Proceedings of the 23rd International Conference on Distributed Computing and Artificial Intelligence (DCAI 2026)*. In press. Naples, Italy: Springer (cited on page 12).

Wang, W. K. et al. (2024). “Addressing wearable sleep tracking inequity: a new dataset and novel methods for a population with sleep disorders”. In: *Proceedings of the Conference on Health, Inference, and Learning*. Vol. 248, pp. 380–396. URL: https://proceedings.mlr.press/v248/wang24a.html (cited on pages 3, 10–12, 56, 57, 59).

Zhou, X., Gao, S., Zeng, P., and Li, L. (2026). “The dose-effect and clinical prediction of the longest apnea duration driving the decrease of blood oxygen: a large sample OSA study with 34-second and 52-second cut-off values was established”. In: *Frontiers in Physiology* 16, p. 1691994. DOI: 10.3389/fphys.2025.1691994 (cited on pages 25, 44, 55).

A. Descriptive Statistics Tables

This appendix presents demographic and clinical descriptive statistics for the $n = 90$ DREAMT participants retained for the duration-based reclassification analysis (Section 4.5.3), complementing the cross-dataset summary already reported in Table 5.1 (Chapter 5).

Table A.1 - Cohort characteristics ($n = 90$).

Variable	N	Mean (SD)	Median (IQR)	Min–Max
Age (years)	90	56.6 (16.7)	58.7 (43.8–70.5)	21.2–86.9
BMI (kg/m^2)	90	33.5 (8.2)	32.0 (29.0–37.0)	18.0–72.0
AHI (events/h)	90	20.0 (24.5)	11.5 (4.0–27.0)	0.0–127.0
OAHI (events/h)	90	17.4 (23.1)	8.5 (2.0–22.8)	0.0–127.0
Mean SpO_2 (%)	90	94.2 (2.7)	95.0 (93.0–96.0)	84.0–99.0
Arousal Index (events/h)	90	36.1 (25.2)	28.0 (20.0–46.8)	3.0–132.0

Table A.2 - Sex distribution.

Sex	N	%
Female	48	53.3
Male	42	46.7

Table A.3 - AHI severity distribution ($n = 90$).

AHI category	N	%
Normal	24	26.7
Mild	25	27.8
Moderate	21	23.3
Severe	20	22.2

B. Goodness-of-Fit Diagnostics for EVT Models

The population-level goodness-of-fit diagnostics for the truncate GPD model (density overlay, log-log CCDF, Q–Q plot) are presented in Figure 4.6 (Chapter 4). This appendix complements that population-level view with the individual-level truncated GPD parameter estimates for the $n = 71$ participants with at least 10 declustered events (Section 4.3.2), of which only the aggregated severity-stratified summary is reported in Table 4.3.

Participant	N	$\hat{\xi}$	$\hat{\sigma}$	P_{90}	P_{95}	P_{99}	AHI	Severity
S002	42	-0.126	14.30	36.06	42.54	55.57	19	Moderate
S003	31	-0.267	19.53	38.99	44.75	54.65	37	Severe
S004	45	-0.158	21.53	48.49	57.58	75.21	99	Severe
S005	61	-0.357	16.31	30.01	33.45	38.80	20	Moderate
S006	18	0.361	0.10	23.30	29.99	53.79	5	Mild
S007	30	-0.317	29.64	53.28	61.22	74.15	15	Moderate
S008	42	-0.051	27.41	68.46	84.76	120.47	48	Severe
S009	45	-0.490	11.58	19.22	20.49	22.21	21	Moderate
S010	38	-0.490	42.87	62.42	69.64	79.38	13	Mild
S011	11	0.398	0.10	25.37	33.51	63.79	2	Normal
S012	24	-0.490	11.77	19.48	20.79	22.55	7	Mild
S013	21	-0.226	11.54	26.68	30.22	36.59	9	Mild
S014	51	-0.024	16.02	45.36	55.64	78.85	21	Moderate
S016	41	-0.490	24.58	37.16	40.90	45.95	34	Severe
S017	41	-0.383	19.54	34.03	37.99	43.98	30	Severe
S018	27	-0.437	21.76	35.24	39.04	44.46	6	Mild
S019	12	0.122	17.30	59.22	76.94	124.34	4	Normal
S020	37	-0.490	35.14	51.74	57.49	65.25	22	Moderate
S022	17	-0.490	27.08	40.62	44.84	50.53	11	Mild

Table B.1 – continued

Participant	N	$\hat{\xi}$	$\hat{\sigma}$	P_{90}	P_{95}	P_{99}	AHI	Severity
S023	17	-0.490	18.53	28.82	31.41	34.90	6	Mild
S025	42	-0.490	24.29	36.76	40.45	45.42	15	Moderate
S026	27	0.059	24.80	72.63	93.21	144.39	24	Moderate
S027	16	-0.385	12.93	23.86	26.13	29.57	8	Mild
S028	58	-0.289	32.13	59.17	68.62	84.46	31	Severe
S030	31	-0.490	46.79	67.82	75.79	86.53	21	Moderate
S031	69	-0.391	31.79	52.33	59.22	69.54	43	Severe
S032	63	-0.348	19.09	34.72	39.03	45.81	22	Moderate
S033	12	-0.490	10.85	18.21	19.34	20.86	2	Normal
S034	12	-0.490	41.80	60.94	67.95	77.42	5	Mild
S035	27	-0.490	52.57	75.80	84.87	97.10	46	Severe
S037	12	0.524	9.11	74.14	114.22	288.48	29	Moderate
S038	48	-0.490	23.73	36.00	39.58	44.41	20	Moderate
S039	21	-0.490	20.37	31.35	34.29	38.26	8	Mild
S040	21	-0.181	20.59	45.33	53.39	68.60	6	Mild
S043	22	-0.442	66.13	99.15	112.47	131.38	13	Mild
S044	24	-0.479	40.34	59.60	66.56	76.06	10	Mild
S045	22	-0.490	29.88	44.49	49.24	55.65	7	Mild
S047	33	-0.490	27.25	40.86	45.11	50.84	76	Severe
S048	22	-0.370	29.15	49.42	56.05	66.21	7	Mild
S050	10	-0.490	35.86	52.73	58.62	66.56	3	Normal
S052	33	-0.185	20.60	45.14	53.10	68.09	6	Mild
S053	38	-0.490	29.81	44.39	49.13	55.52	20	Moderate
S054	41	-0.484	32.44	48.34	53.67	60.92	24	Moderate
S056	19	-0.260	16.83	34.63	39.59	48.16	8	Mild
S057	15	0.433	2.10	35.42	49.53	104.40	3	Normal
S058	29	-0.490	30.62	45.51	50.40	57.00	64	Severe
S059	18	-0.490	28.58	42.69	47.19	53.27	127	Severe
S061	35	-0.490	67.28	96.11	107.97	123.97	37	Severe
S063	10	-0.490	11.75	19.46	20.76	22.52	2	Normal
S064	19	-0.490	23.00	34.99	38.43	43.07	37	Severe
S067	10	0.197	17.50	66.74	89.51	156.08	2	Normal

Table B.1 – continued

Participant	N	$\hat{\xi}$	$\hat{\sigma}$	P_{90}	P_{95}	P_{99}	AHI	Severity
S068	25	-0.490	19.75	30.49	33.32	37.12	30	Severe
S069	31	0.183	12.39	50.71	66.72	112.73	19	Moderate
S070	59	-0.369	37.78	62.91	71.83	85.55	33	Severe
S071	44	-0.490	27.23	40.82	45.07	50.80	70	Severe
S072	18	-0.490	39.04	57.13	63.62	72.38	69	Severe
S073	37	-0.235	24.59	49.52	57.79	72.51	14	Mild
S074	15	-0.490	42.92	62.49	69.72	79.48	5	Mild
S077	22	-0.300	19.24	37.01	42.10	50.55	6	Mild
S078	26	0.049	0.10	11.43	11.89	13.03	6	Mild
S079	62	-0.366	14.47	26.83	29.68	34.08	29	Moderate
S080	25	-0.171	29.63	63.16	75.48	99.03	12	Mild
S081	25	-0.219	32.09	64.06	75.67	96.71	18	Moderate
S083	25	-0.490	29.42	43.85	48.51	54.80	50	Severe
S086	12	-0.490	21.14	32.41	35.50	39.66	4	Normal
S088	39	0.120	6.53	30.50	37.86	57.51	28	Moderate
S089	35	-0.087	10.49	30.07	35.37	46.50	33	Severe
S091	15	-0.313	26.34	48.07	55.10	66.58	10	Mild
S092	48	-0.490	35.47	52.20	58.02	65.86	21	Moderate
S093	49	-0.063	12.67	35.79	42.84	58.08	17	Moderate
S098	52	-0.164	20.44	46.03	54.47	70.72	23	Moderate

Note: $\hat{\xi} = -0.490$ indicates the lower bound imposed during numerical optimization was active for that participant (a near-bounded fit, not a free estimate).

C. Reclassification Table — Combined Classifier

Table 4.5 (Chapter 4) reports the aggregated 4×4 reclassification matrix between AHI-only severity (A_i) and the combined classifier ($C_i = \max(A_i, B_i)$). This appendix lists the $n = 24$ participants whose severity category changed under the combined classifier, together with the maximum event duration (D^{max}) that drove the reclassification.

Table C.1 - Participants reclassified by the combined classifier ($T_1 = 34\,\text{s}$, $T_2 = 52\,\text{s}$).

Participant	AHI	D^{max} (s)	AHI-only severity	Combined severity
S075	0	37.5	Normal	Mild
S065	1	94.1	Normal	Moderate
S085	1	44.5	Normal	Mild
S051	1	45.4	Normal	Mild
S011	2	37.8	Normal	Mild
S076	2	48.9	Normal	Mild
S067	2	98.8	Normal	Moderate
S050	3	55.3	Normal	Moderate
S057	3	53.5	Normal	Moderate
S086	4	38.4	Normal	Mild
S049	4	67.8	Normal	Moderate
S019	4	88.4	Normal	Moderate
S034	5	67.2	Mild	Moderate
S074	5	72.2	Mild	Moderate
S052	6	69.7	Mild	Moderate
S040	6	68.8	Mild	Moderate
S045	7	52.9	Mild	Moderate
S048	7	63.8	Mild	Moderate
S044	10	76.0	Mild	Moderate
S091	10	64.1	Mild	Moderate
S080	12	104.7	Mild	Moderate
S043	13	131.6	Mild	Moderate
S010	13	77.5	Mild	Moderate
S073	14	76.5	Mild	Moderate

Twelve of the 24 reclassified participants had Normal AHI but a clinically relevant maximum event duration, directly supporting the undetected-risk finding discussed in Section 4.5.4.

Table C.2 - Severity distribution before and after combination ($n = 90$).

Severity	AHI alone	Combined
Normal	24	12
Mild	25	19
Moderate	21	39
Severe	20	20