UNCERTAINTIES IN HYDROLOGICAL DATA

UNCERTANTIES IN DATA

Did you know uncertainties in hydrological data are a systematic problem that often lead users to treat hydrological data as error-free information? A recent evaluation found that up to 50% of streamflow estimates at more than 1,500 site-months were unreliable (Westerberg et al. 2024, p. 1). Errors persist due to limited resources, inadequate protocol assurance, and seasonal factors. To address this, researchers advocate using "soft information", qualitative observations, field knowledge, and contextual factors affecting data quality that are not captured by measurements or rating curves. These elements form the basis of a perceptual model of uncertainty, helping explain why and where uncertainties occur within a dataset.

SOFT INFORMATION IN HYDROLOGICAL DATA 101

Soft information helps explain uncertainty arising from incomplete knowledge of a gauging station’s condition. For perceptual models, it is commonly grouped into three categories: **station characteristics, climate and flow regime, and catchment characteristics.**

1. STATION CHARACTERISTICS

Factors such as site remoteness, unstable riverbeds, and changing gauging techniques can affect data quality. Remote stations are harder to access during high-flow events, increasing uncertainty in peak flow estimates. Vegetation growth can raise water levels, while floodplain overflow may alter high-flow rating curves. Older records often rely on manual observations, whereas modern records use automated systems.

2. CLIMATE AND FLOW REGIME

Predictable flow patterns, such as snowmelt-driven floods, are generally easier to monitor than rainfall-driven events. Ice-covered rivers present additional challenges because discharge often cannot be measured directly and must be estimated over extended periods.

3. CATCHMENT CHARACTERISTICS

Catchment geography influences uncertainty. Large catchments, lakes, and absorbent soils slow water movement and stabilise flows, making measurements more reliable. In contrast, steep or mountainous catchments often experience rapid flow changes and unstable riverbeds, increasing uncertainty.

SOFT INFORMATION AS A TOOL FOR UNCERTAINTY ANALYSIS

Soft information is essential for developing perceptual models of uncertainty. Station photographs, aerial imagery, and site visits help identify factors affecting hydraulic control and changes in the stage-discharge relationship. Information on soils, riverbed materials, vegetation, and gauging structures can indicate erosion, sedimentation, and channel change that may increase uncertainty in rating curve models.

Soft information is also valuable during ice conditions and high-flow events. The timing, duration, and predictability of peak flows influence measurement accuracy. In small catchments with rapid rainfall-runoff responses, short flood peaks are often difficult to gauge, especially during nighttime or winter conditions.

CASE IN MYLLYMÄKI

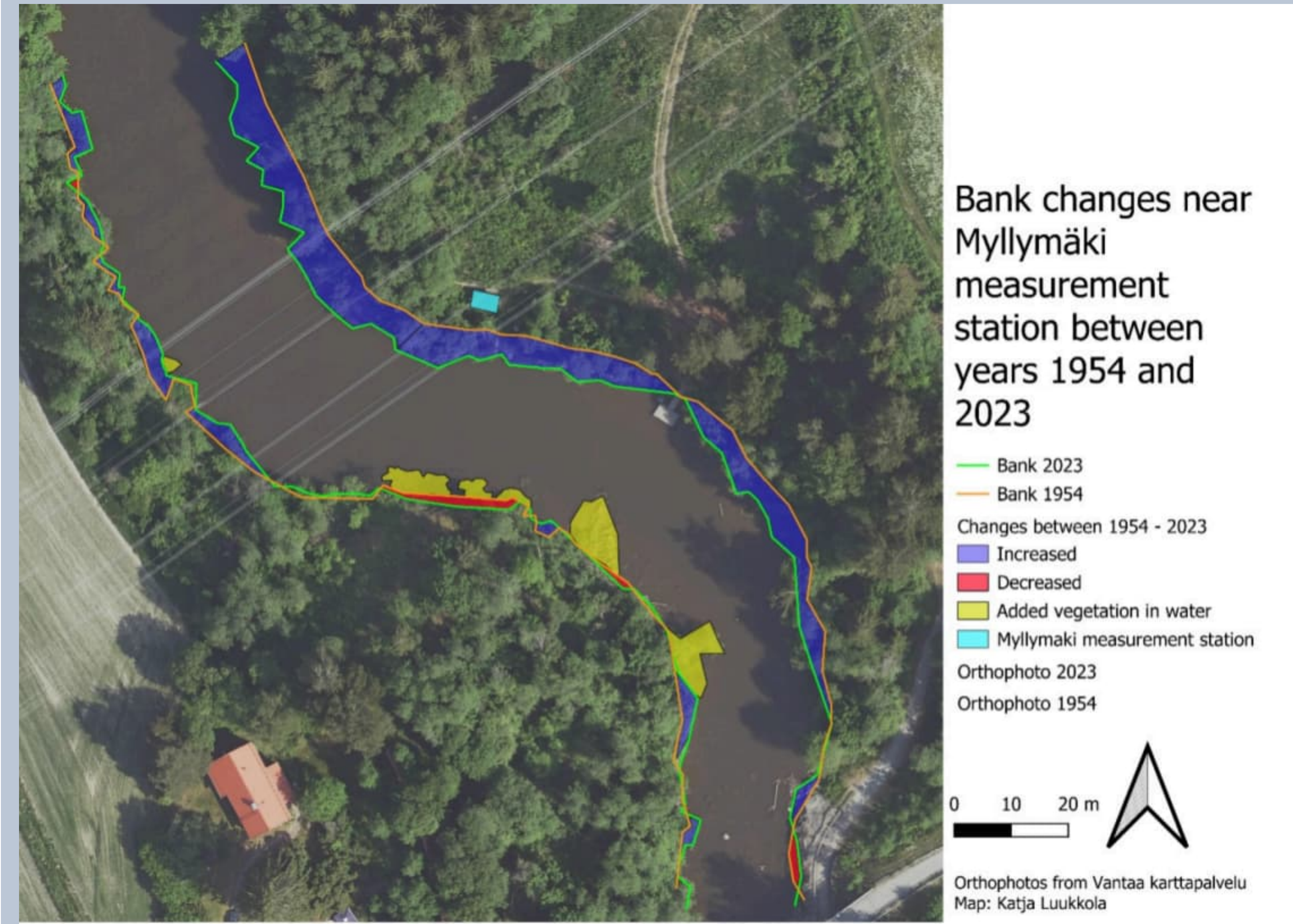


Figure 1 Bank changes near Myllymäki measurement station between years 1954 and 2023.

To evaluate soft uncertainties, a spatial vegetation analysis was conducted by comparing aerial orthophotos of the Myllymäki channel from 1954 and 2023. The results are shown in Figure 1, where the analysis was performed by overlapping the two images and identifying differences. The dominant feature is the channel narrowing over time, shown in **blue**. Some erosion (**red**) and increased vegetation within the channel (**yellow**) are also visible.

Overall, the river channel has narrowed considerably, but quantifying the error caused by 70 years of vegetation growth is impossible. Table 1 shows water level and discharge measurements from 1959 and 2023. The observed differences may be partly explained by vegetation growth since 1954, which likely reduced flow velocity and discharge capacity while increasing water levels.

CASE IN MYLLYMÄKI

TABLE 1 WATER LEVELS AND DISCHARGE: 1959 VS 2023

The core of the data problem is best seen from the station's rating curves shown in Figures 2 and 3. The rating curve method has been used by SYKE to translate water level into flow volume. In an ideal situation, the blue observation points would align with the black line. Figure 3 shows that data from 1985 onwards follows this pattern closely, whereas data from 1959 to 1984 consistently deviates from the theoretical baseline. During this early period, manual corrections were apparently applied to the data, but they were never documented. This lack of communication affects both open-data users and SYKE staff. If left uncommunicated, this early dataset remains unreliable for climate modelling.

Month	Water Level (1959)	Water Level (2023)	Discharge (1959)	Discharge (2023)
May	2348.06	2365.39	8.1	8.03
July	2323.39	2325.84	3.04	2.67

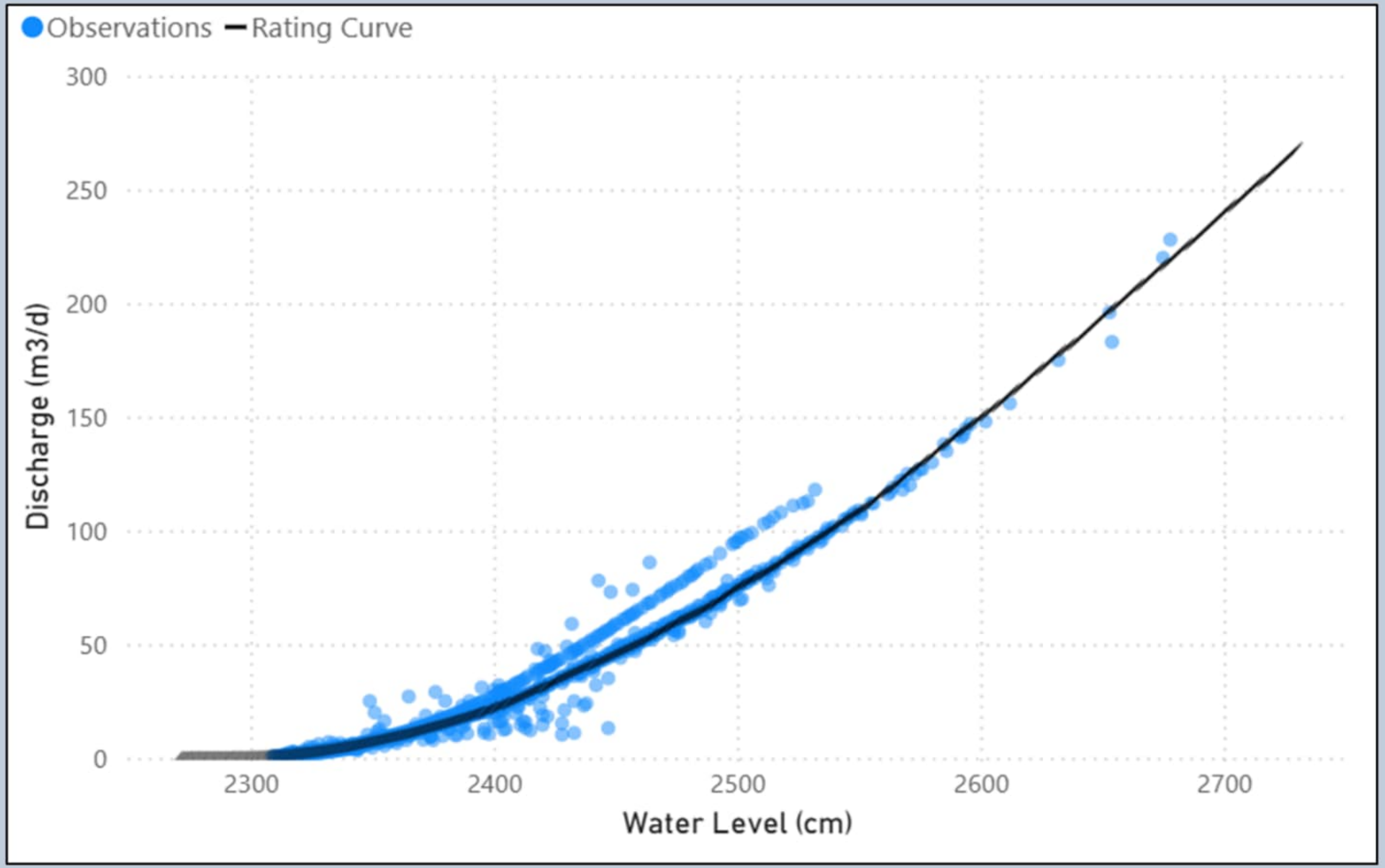


Figure 2 The rating curve which translates water level into flow volume. From 1959 to 1984.

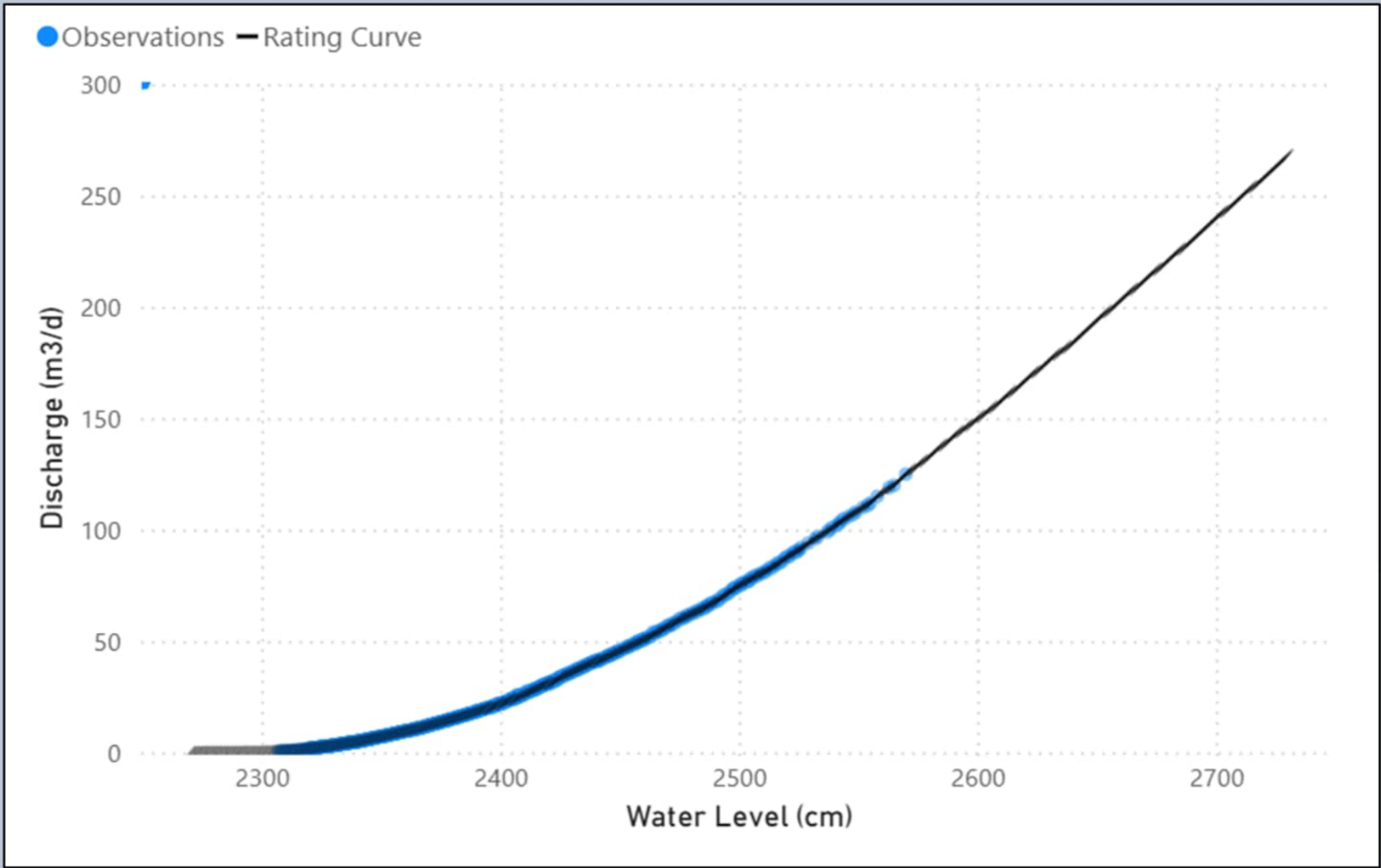


Figure 3 The rating curve which translates water level into flow volume. From 1985 to 2026.

To demonstrate the vulnerabilities better, a tier ranking system was created for SYKE's water level and discharge data. This system translates complex environmental events, such as summer weed growth or undocumented rating curve shifts, into clear text-based warnings. The proposed framework tags every data point based on its risk level using three tiers:

Tier 1 – Low Risk	Tier 2 – Medium Risk	Tier 3 – High risk
• Stable riverbed & open water • Modern digital technology era • No quality errors	Same as Tier 1, but: • Manual staff gauge readings • Interpolated/ extrapolated values • Reconstructed historical data & minor operational variance	• Bed instability • Severe boundary contamination • Presence of ice • Critical database quality notes & severe operational flaws

Figure 4 Definitions of the Tiers 1-3

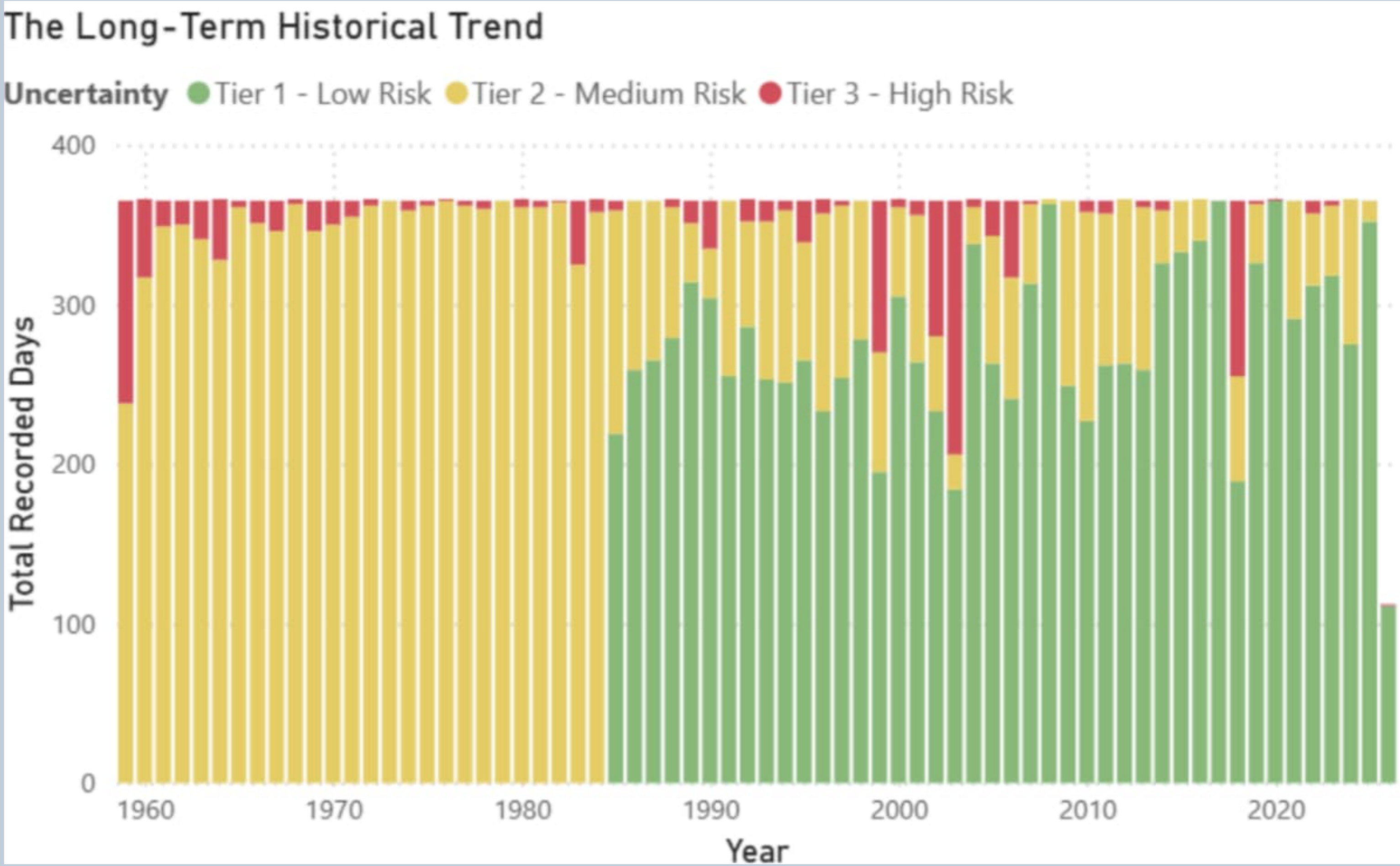
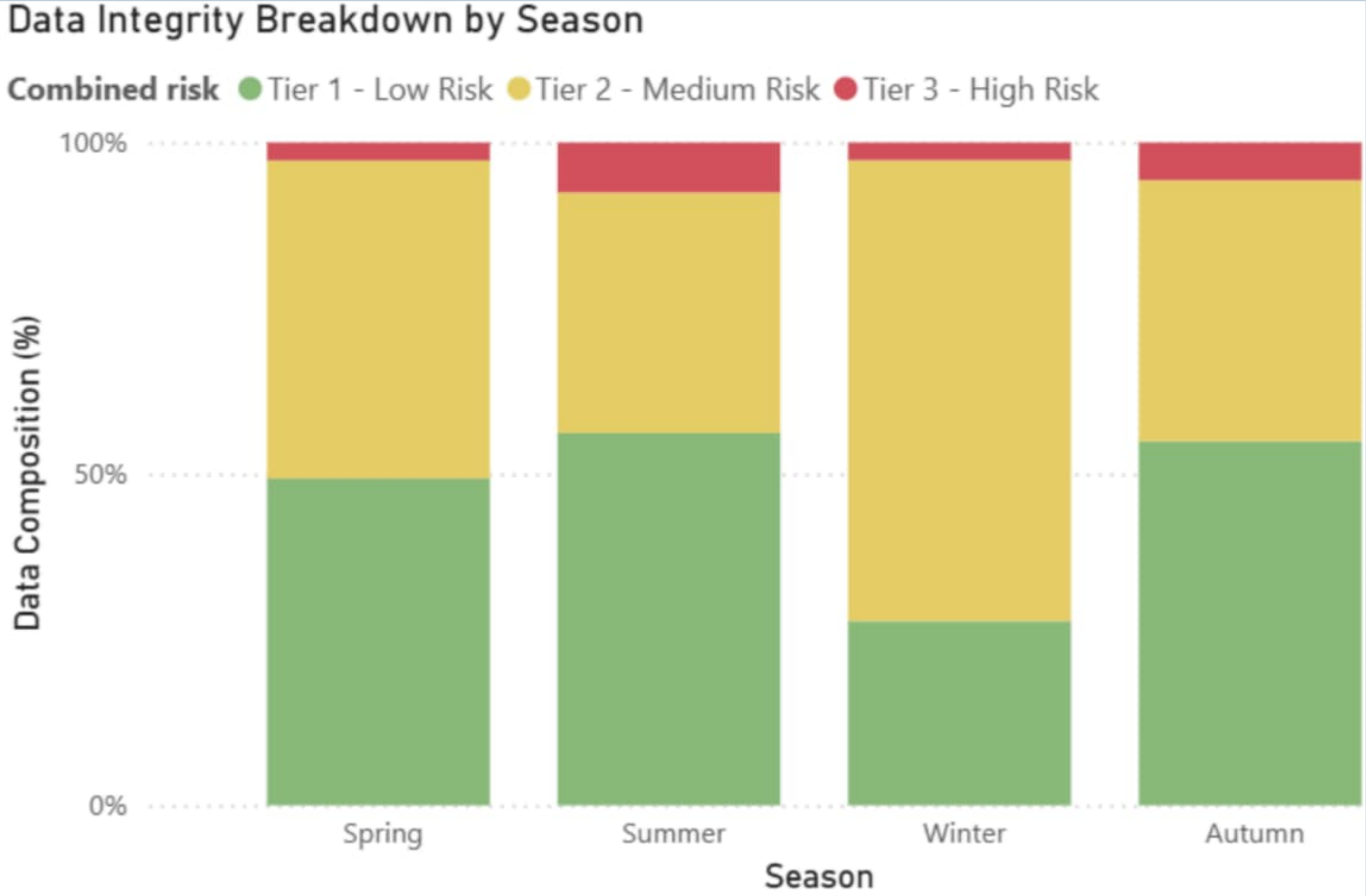


Figure 5 Long term Historical trend of SYKE data reliability divided into the three Tiers

Figure 5 gives us a quick health check of the station's data over time. The figures were created using Excel for calculations and Power BI for visualisation. Overall, the dataset consists of 48% Tier 1, 49% Tier 2, and 2% **Tier 3** data. By combining Tier 1 data with the reliable portion of Tier 2 values, we calculated an overall Data Integrity Score of 71%, representing the share of data considered safe to use.

One pattern stands out immediately: Tier 2 dominates the period before 1985. This highlights the undocumented manual era of the rating curve, where uncertainty is much higher. While data quality generally improves in the modern era, it still fluctuates, reminding us that newer technology does not automatically guarantee better data.



The seasonal chart in Figure 6 reveals a clear environmental pattern. The summer and autumn months show the clearest impact of vegetation growth, which can block the channel and artificially raise water level readings. As a result, the number of **Tier 3** red flags increases during the warmer months.

This pattern demonstrates how natural environmental changes can directly influence data quality. Even when measurement systems remain unchanged, seasonal conditions can introduce additional uncertainty that users should be aware of when interpreting the data.

Figure 5 Long term Historical trend of SYKE data reliability divided into the three Tiers

The years 2003 and 2018 stand out in Figure 5, demonstrating that modern sensor data is not inherently flawless. Both occur within the automated measurement era yet show the highest concentrations of Tier 3 risk. These issues were caused by extreme drought conditions in 2003 and an uncommunicated four-month sensor malfunction in 2018.

Because such problems are often hidden in standard download portals, researchers may assume the data is reliable. Our framework isolates these anomalies while preserving the validity of the surrounding decades

WHY IS THIS IMPORTANT?

One impact of climate change on river morphology is increased vegetation growth on the riverbed due to higher water temperatures, which can artificially raise water levels. Although measured water levels may increase, this does not necessarily mean that more water is flowing through the channel, causing streamflow to be underestimated. Inaccurate streamflow estimates can affect a wide range of SYKE data users, including students, companies, government authorities, and private individuals who rely on the data for research, planning, fishing, and managing summer cottages. If uncertainties are not well communicated, users risk making decisions based on inaccurate information.

Having an overview of uncertainties in the data, understanding the available soft information, and developing perceptual models that incorporate this information can help mitigate these risks. This supports a functioning environmental data system where users can better assess data quality and reliability.

RESOURCES

Blogpost is written based on the research done by Wesberg et al 2024 :
DOI: 10.1002/hyp.15145

SYKE GROUP:
BIANCA CLEPCEA
KERTTU VAINIOMÄKI
KATJA LUUKKOLA
SATU JORONEN
LUMI HULTKRANTZ

MENTOR FROM SYKE:
SAMI SOOSALU