Rainfall, River Flow and Storm Overflows in Devon and Cornwall: Project Report

Oct 19, 2026 · Soroush Azarinejadianlangroudi

Summary

In the South West, river flow depends more on the last two weeks of rain than on the rain that fell today, and storm-overflow activity in 2025 follows rainfall closely but not completely.

Using open Environment Agency and South West Water data for 2015–2025, I found that 14-day rainfall correlates 0.77 with log river flow on the Exe, against 0.34 for same-day rain. A simple next-day flow model beats persistence on held-out data (Nash-Sutcliffe efficiency 0.85 against 0.78). In 2025, days with at least 10 mm of rain over the previous two days averaged about 1,580 storm-overflow events, against about 250 on other days (6.4 times as many). The worst 10% of overflow sites account for about 47% of all discharge hours.

The work is reproducible end to end: a Python notebook, 11 SQL queries and a six-page Power BI dashboard, all in one repository: github.com/su7ri/water-rainfall-river-overflow.

Question and data

The question: how does rainfall drive river flow in the South West, and how closely do storm-overflow discharges follow rainfall? Both datasets are open under the Open Government Licence v3.

Source	Content	Coverage
Environment Agency Hydrology API	Daily rainfall at 7 gauges within 60 km of Exeter; daily mean flow on the River Exe (Thorverton) and River Fal (Trenowth)	2015–2025, about 36,000 daily readings

Source	Content	Coverage
Event Duration Monitoring: Storm Overflow Start/Stop Detailed Data	South West Water 2025 storm-overflow events, one row per event	209,711 events at 1,179 sites

Method

The analysis runs in three stages, in this order: Python, SQL, Power BI.

- **Cleaning and checks (Python).** Readings are loaded from the API one series at a time. Missing values are set to NaN, never zero. Suspect, Estimated and Unchecked readings (10.7% of the total) are kept and flagged. The rainfall index is the mean of the 7 gauges with Suspect readings excluded.
- **Rain to flow (Python).** Lagged correlations between rain and flow, then a Ridge regression on log-flow using lagged flow, 3-, 7- and 30-day rainfall and seasonality. It is trained on data up to 2022 and tested on 2023–2024, and compared with a persistence baseline using Nash-Sutcliffe efficiency (NSE) and RMSE.
- **Overflows (Python).** Events are aggregated per day and compared with 2-day rainfall; days are split into wet (at least 10 mm over the previous two days) and other days. Events longer than 7 days are flagged and excluded.
- **SQL (SQLite).** The cleaned tables are loaded into a star schema (`dim_date`, `dim_station`, `fact_daily_readings`, `fact_edm_events`, `fact_edm_daily`) and queried with 11 queries using window functions, `PERCENT_RANK` and `NTILE`: quality checks, rolling rainfall against flow, low-flow days, concentration of discharge by site and day-level spikes.
- **Power BI.** The same tables feed a six-page report with DAX measures, kept in a dedicated measures table, so that the dashboard figures tie to the notebook.

Findings

1. **The rain gauges agree.** Each correlates 0.86–0.91 with the mean of the others, so a regional mean is a reasonable rainfall index.
2. **The Exe responds to rain within a day, but remembers it for weeks.** Flow peaks at a 1-day lag (correlation 0.54). The last 14 days of rain correlate 0.77 with log-flow, against 0.34 for the same day alone.

3. **A simple model beats persistence.** On 2023–2024 data, next-day flow has NSE 0.85 against 0.78 and RMSE 9.5 against 11.5 m³/s. It still overshoots some flood peaks, by +4.5 m³/s on average on the top 5% of flow days.
4. **Overflow activity follows rainfall.** In 2025, days with at least 10 mm of rain over the previous two days averaged about 1,580 overflow events per day, against about 250 on other days (Spearman 0.65).
5. **Rainfall is not the whole story.** December 2025 had roughly 72,000 events against about 39,000 in January, with almost the same rainfall, in a year that was not unusually wet.
6. **Low flow clusters in a few years.** Days below the long-run 10th-percentile Exe flow (2.71 m³/s): 2022 had 133, 2018 had 81 and 2025 had 65, against 6–33 in every other year. Low-flow frequency matters for abstraction and supply planning.
7. **Overflow impact is concentrated.** The worst 10% of sites account for about 47% of all discharge hours. The largest single site is Delabole wastewater treatment works, with about 2,900 hours.

Dashboard

The Power BI report has six pages. The `.pbix` file and a PDF export are in the `powerbi/` folder of the repository.

Page	What it shows
Overview	Annual rainfall, Exe and Fal flow by month, headline figures, year and season filters
Rain to River	Daily rain and Exe flow for a chosen water year; 14-day rainfall against flow with a trend line
Low Flow	Days below the 10th-percentile flow by year, and the threshold itself (2.71 m³/s)
Storm Overflows 2025	Events against 2-day rain, wet-day against dry-day averages (ratio 6.4), top 10 sites by discharge hours
Data Quality	Readings by quality class and station, share not rated Good (10.7%), long events excluded (356)
Method and Limits	Sources, cleaning decisions and limitations

Data quality and limitations

What I found in the data and how I handled it:

- The Fal at Trenowth had two readings per day for one year. I kept the 09:00 reading, as at every other station.
- 10.7% of daily readings are Suspect, Estimated or Unchecked. They are kept and flagged. Hemyock has the highest share (about 23%), followed by Werrington Park (18%) and Ashcombe (16%).
- 356 overflow events last more than 7 days (up to about 97 days), likely monitor faults or unclosed events. They are flagged and excluded from counts.
- 23 overflow monitors have no site name in the source file. They are kept and shown by their monitor ID.

Limitations to keep in mind:

- Correlation is not causation, and only one flow catchment (the Exe) is modelled in detail.
- The overflow data is self-reported by the water company, has not been independently verified, and covers a single year.
- Event counts are not discharge volumes: a long, small spill and a short, large one count the same.
- Sites are ranked by monitor ID, so one treatment works with several monitors can appear more than once.
- The flow model overshoots some flood peaks.

Relevance to water resources and next steps

The project covers the same ground as day-to-day water-resources reporting: turning rainfall and flow records into checked, documented figures, flagging data-quality problems before they reach a result, and presenting the outcome in a form non-specialists can use. The low-flow and rain-to-flow results speak to abstraction and supply planning; the overflow results show where an investigation would pay off most.

Next steps, in order of value:

1. Add more catchments and several years of overflow data, so that wet-year and dry-year behaviour can be compared.

2. Use discharge duration per site as the impact measure and link sites to their receiving waters.
3. Try gradient-boosted and probabilistic flow forecasts, with uncertainty bands for flood peaks.

Contains public sector information licensed under the Open Government Licence v3.0.