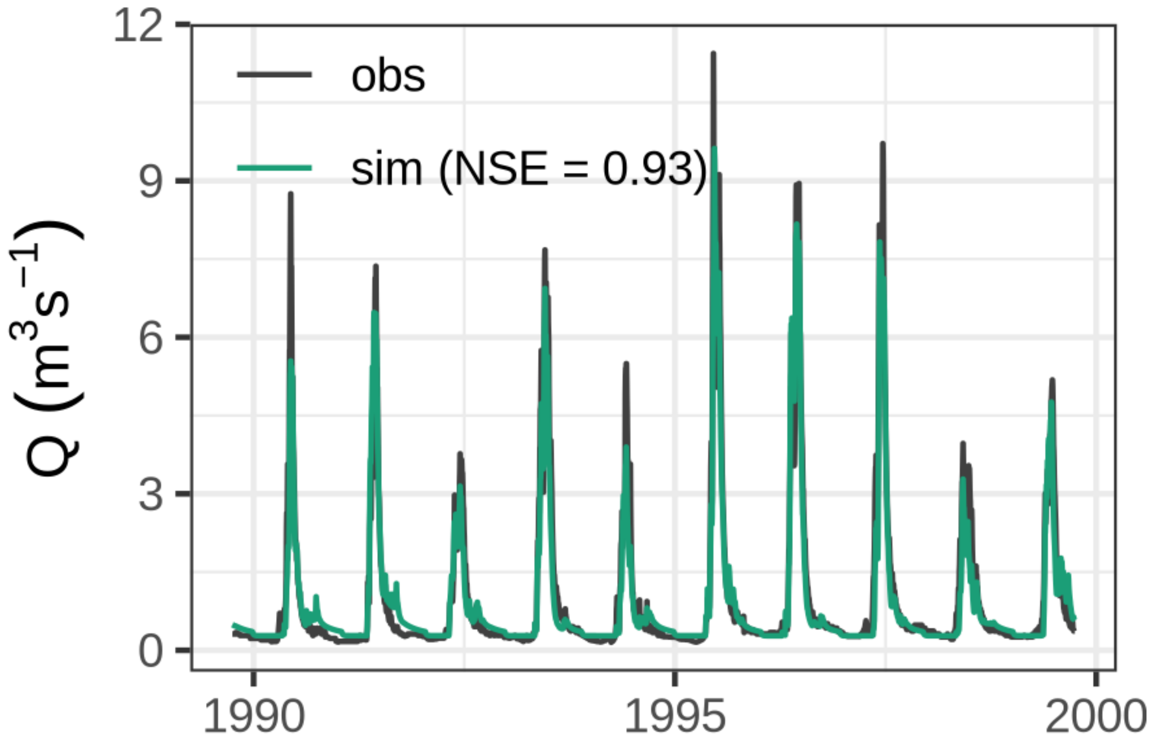
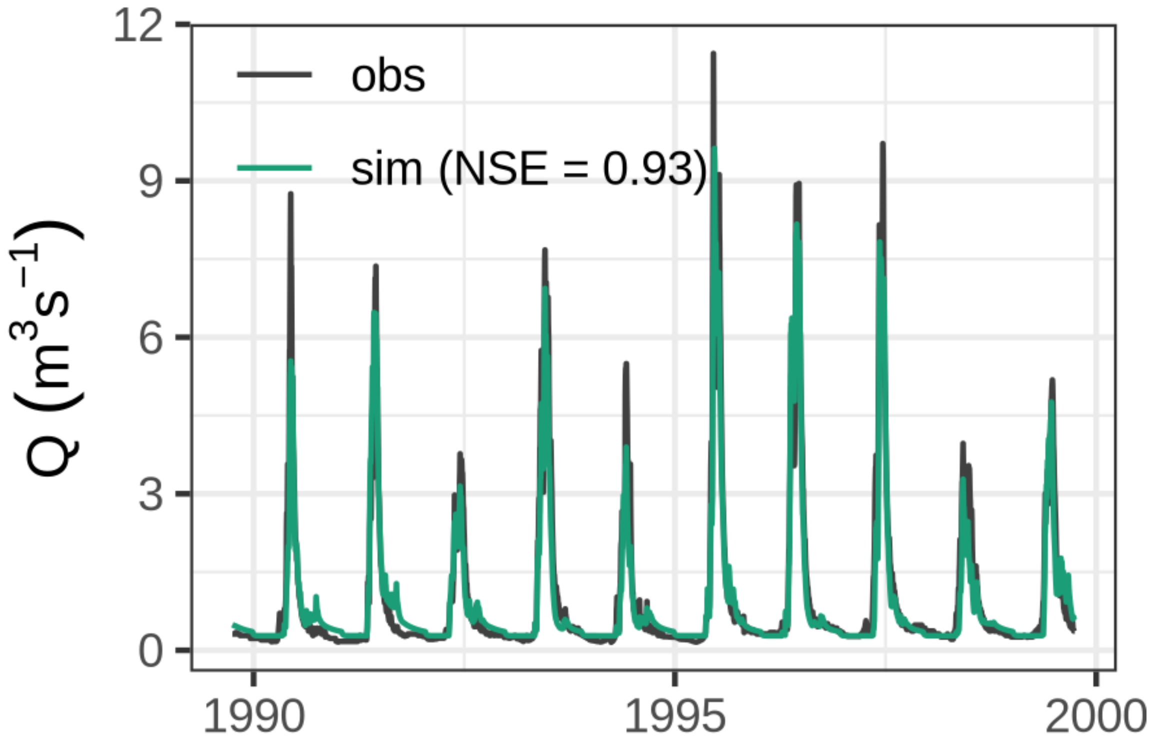
Does a high NSE always indicate a good model?



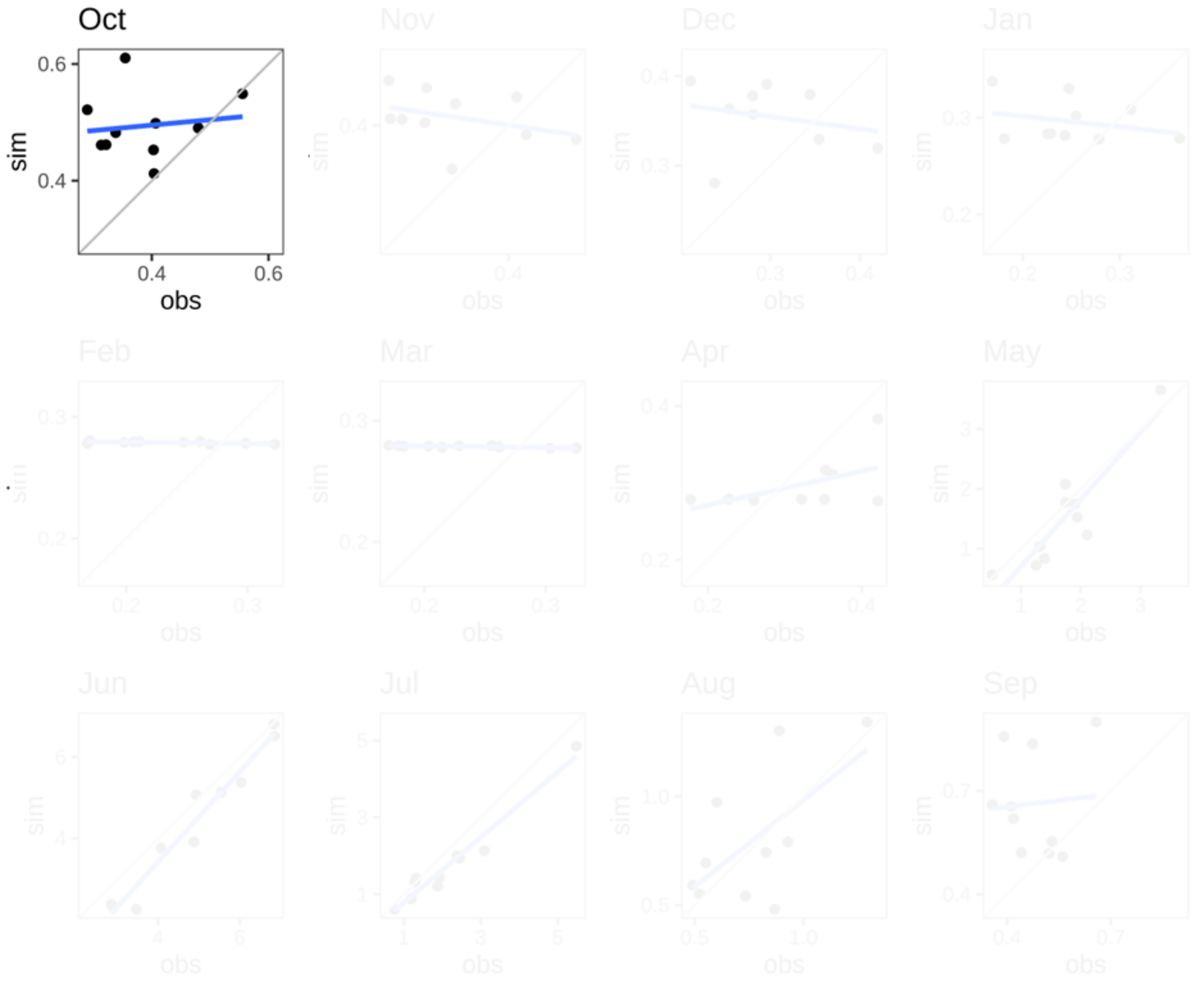
Sacha Ruzzante[1], Wouter Knoben[2], Thorsten Wagener[3], Tom Gleeson[1], and Markus Schnorbus[1]

[1]University of Victoria, Canada
[2]University of Calgary, Canada
[3]University of Potsdam, Germany

Does a high NSE always indicate a good model?



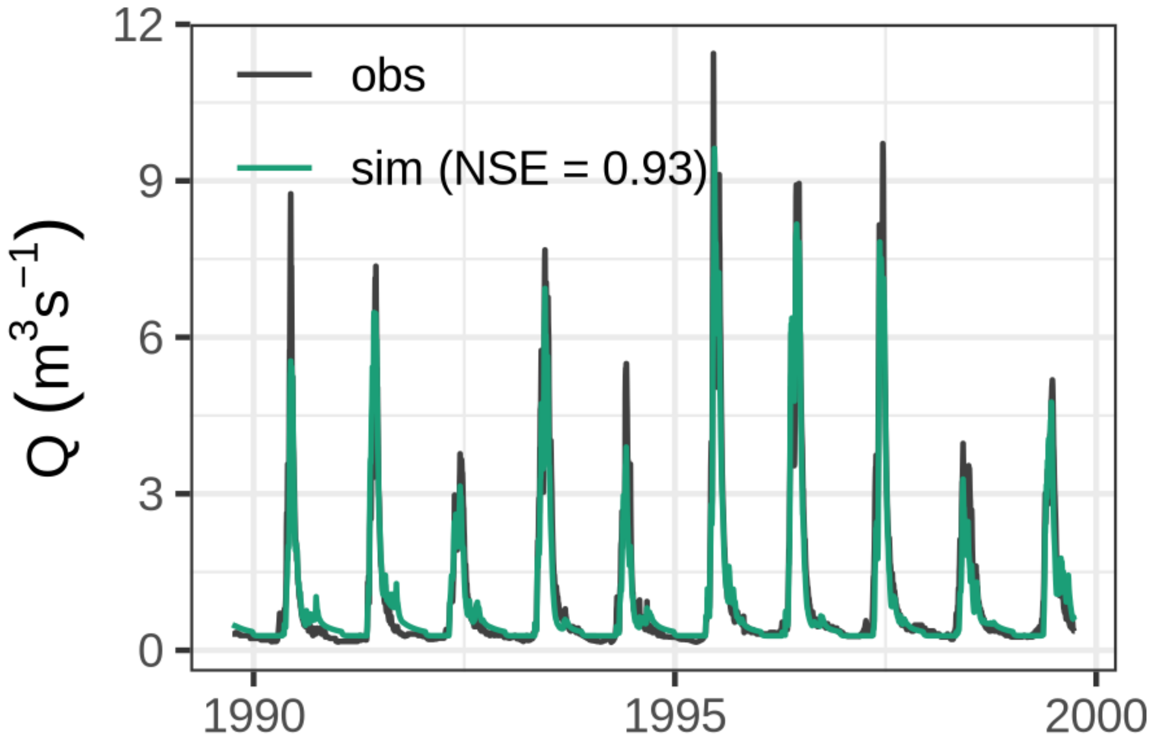
… maybe not



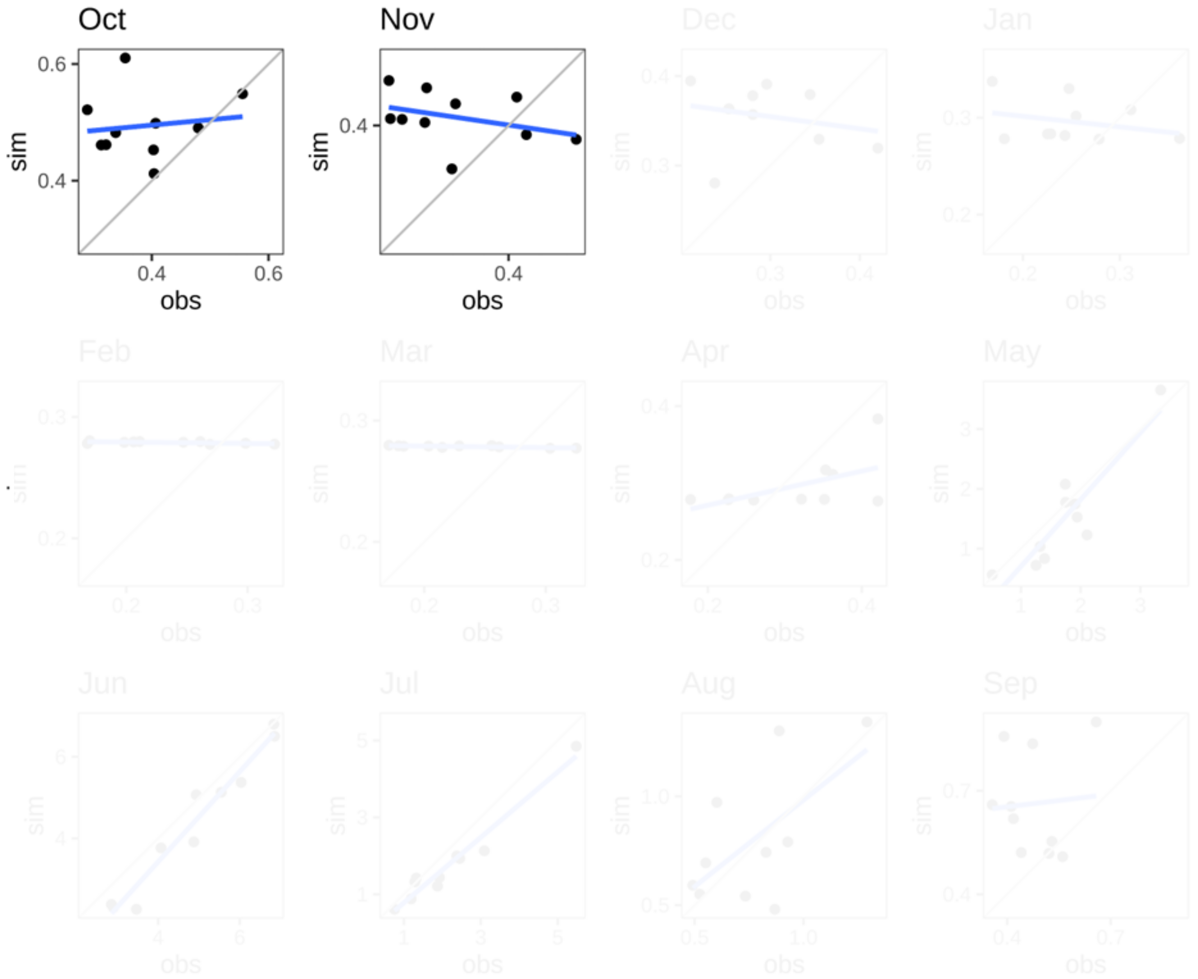
Sacha Ruzzante[1], Wouter Knoben[2], Thorsten Wagener[3], Tom Gleeson[1], and Markus Schnorbus[1]

[1]University of Victoria, Canada
[2]University of Calgary, Canada
[3]University of Potsdam, Germany

Does a high NSE always indicate a good model?



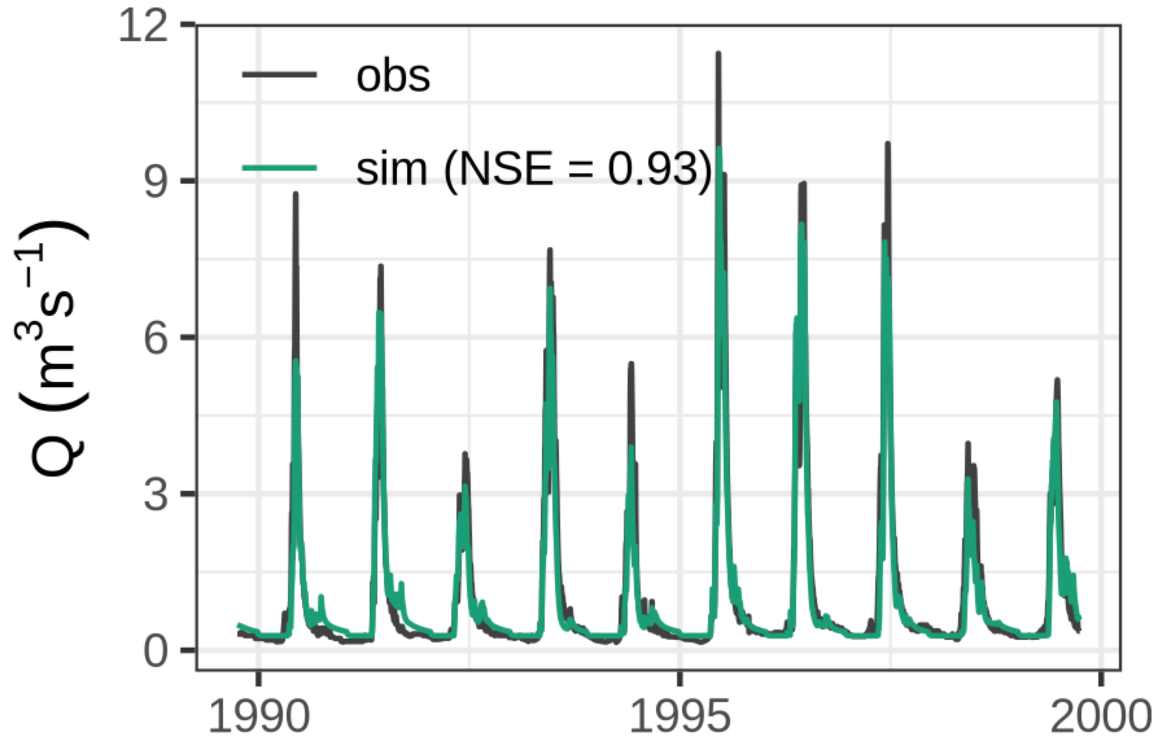
… maybe not



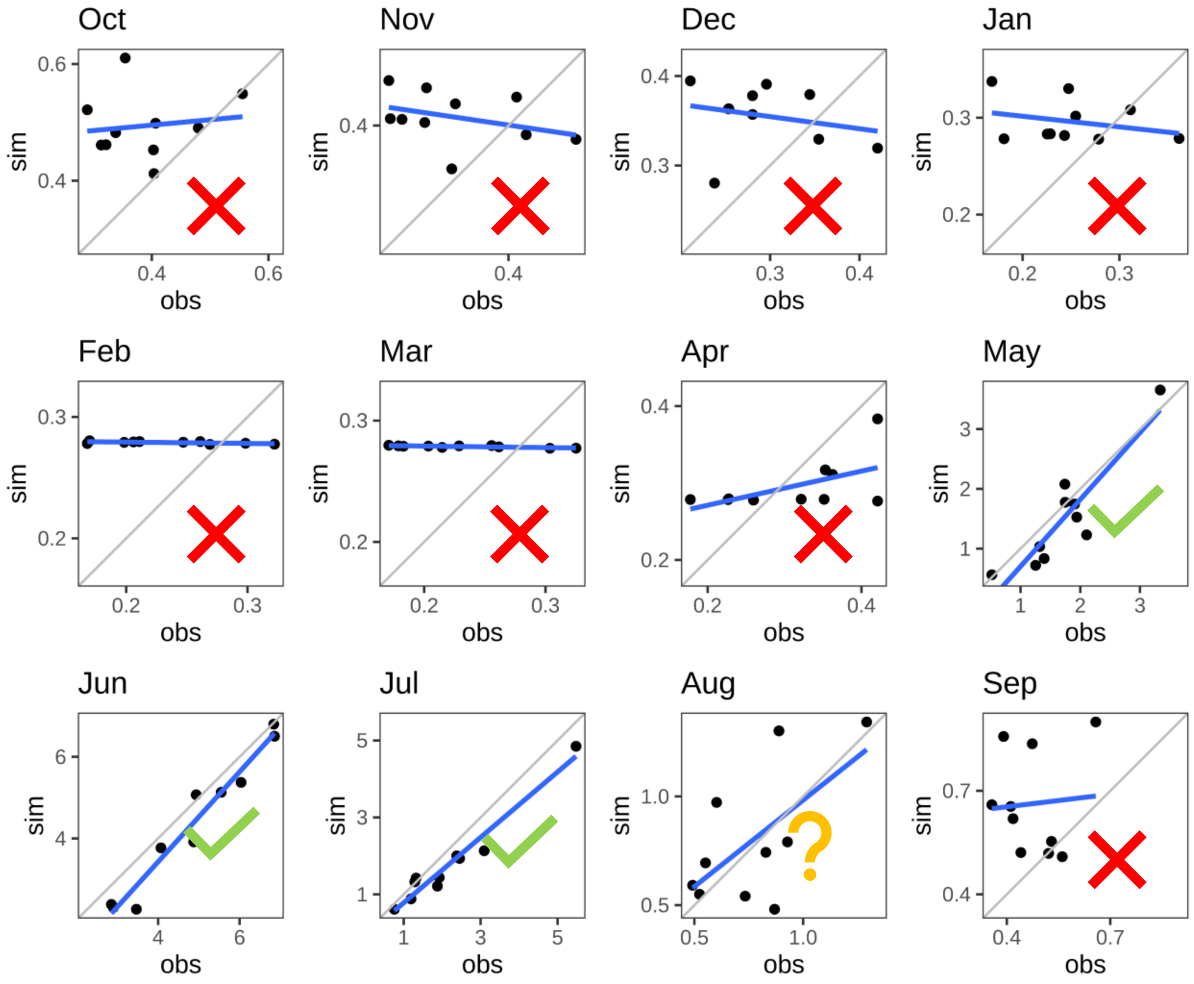
Sacha Ruzzante[1], Wouter Knoben[2], Thorsten Wagener[3], Tom Gleeson[1], and Markus Schnorbus[1]

[1]University of Victoria, Canada
[2]University of Calgary, Canada
[3]University of Potsdam, Germany

Does a high NSE always indicate a good model?



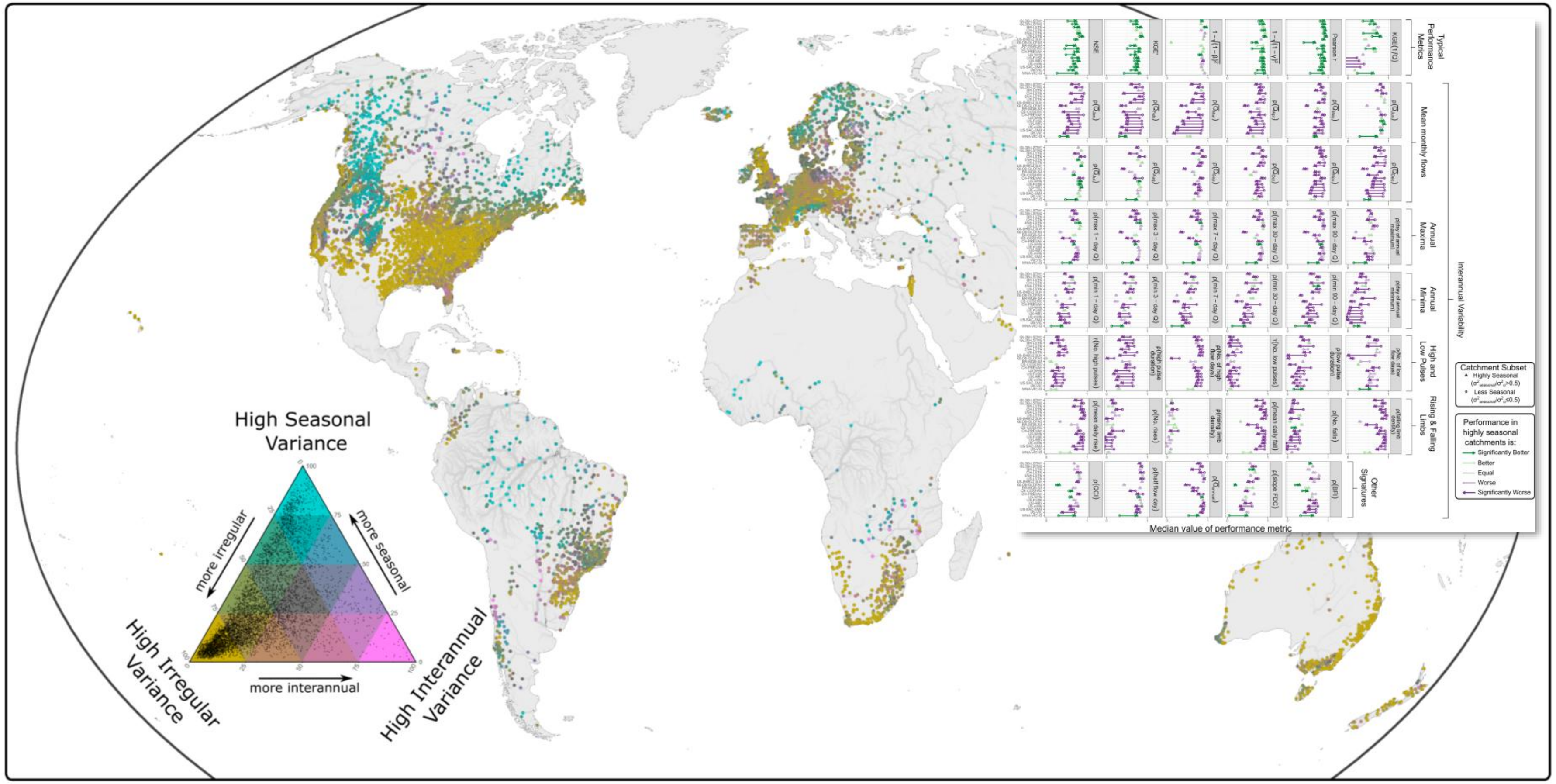
… maybe not



Sacha Ruzzante[1], Wouter Knoben[2], Thorsten Wagener[3], Tom Gleeson[1], and Markus Schnorbus[1]

[1]University of Victoria, Canada
[2]University of Calgary, Canada
[3]University of Potsdam, Germany

Hydrologic models achieve high NSE in seasonal catchments but are **worse at simulating interannual variability**



1) LSTM: Kratzert et al. (2024)
2) LSTM: Arsenault et al. (2022)
3) LSTM: Ruzzante et al. (2026)
4) LSTM: Kraft et al. (2025)
5) LSTM: Yang et al. (2025)
6) HBV: Seibert et al. (2018)
7) PREVAH: Kraft et al. (2025)
8) FUSE: Addor; Kratzert. (2019)
9) VIC-GL: Schnorbus (2018)
10) δHBV: Song et al. (2025)
11) LSTM: Nearing et al. (2024)
12) COSERO: Klingler et al. (2021)
13) SAC-SMA: Newman et al. (2017)
14) mHM: Mizukami et al. (2017)
15) VIC: Newman et al. (2017)
16) NHM: Regan et al. (2019)
17) MGB: Siquiera et al. (2018)
18) GloFAS: Nearing et al. (2024)

Sacha Ruzzante[1], Wouter Knoben[2], Thorsten Wagener[3], Tom Gleeson[1], and Markus Schnorbus[1]
[1]University of Victoria, Canada
[2]University of Calgary, Canada
[3]University of Potsdam, Germany

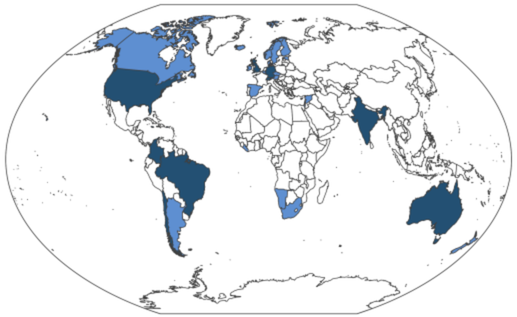
ALSO:

Caravan-CMIP6

CMIP
Coupled Model Intercomparison Project

Bias-corrected multi-model ensemble climate projections for:

Caravan
CAMELS
CAMELS-AUS-v2
CAMELS-BR
CAMELS-CH
CAMELS-CL
CAMELS-COL
CAMELS-DE
CAMELS-GB-v2
CAMELS-IND



Sacha Ruzzante[1], Wouter Knoben[2], Thorsten Wagener[3], Tom Gleeson[1], and Markus Schnorbus[1]

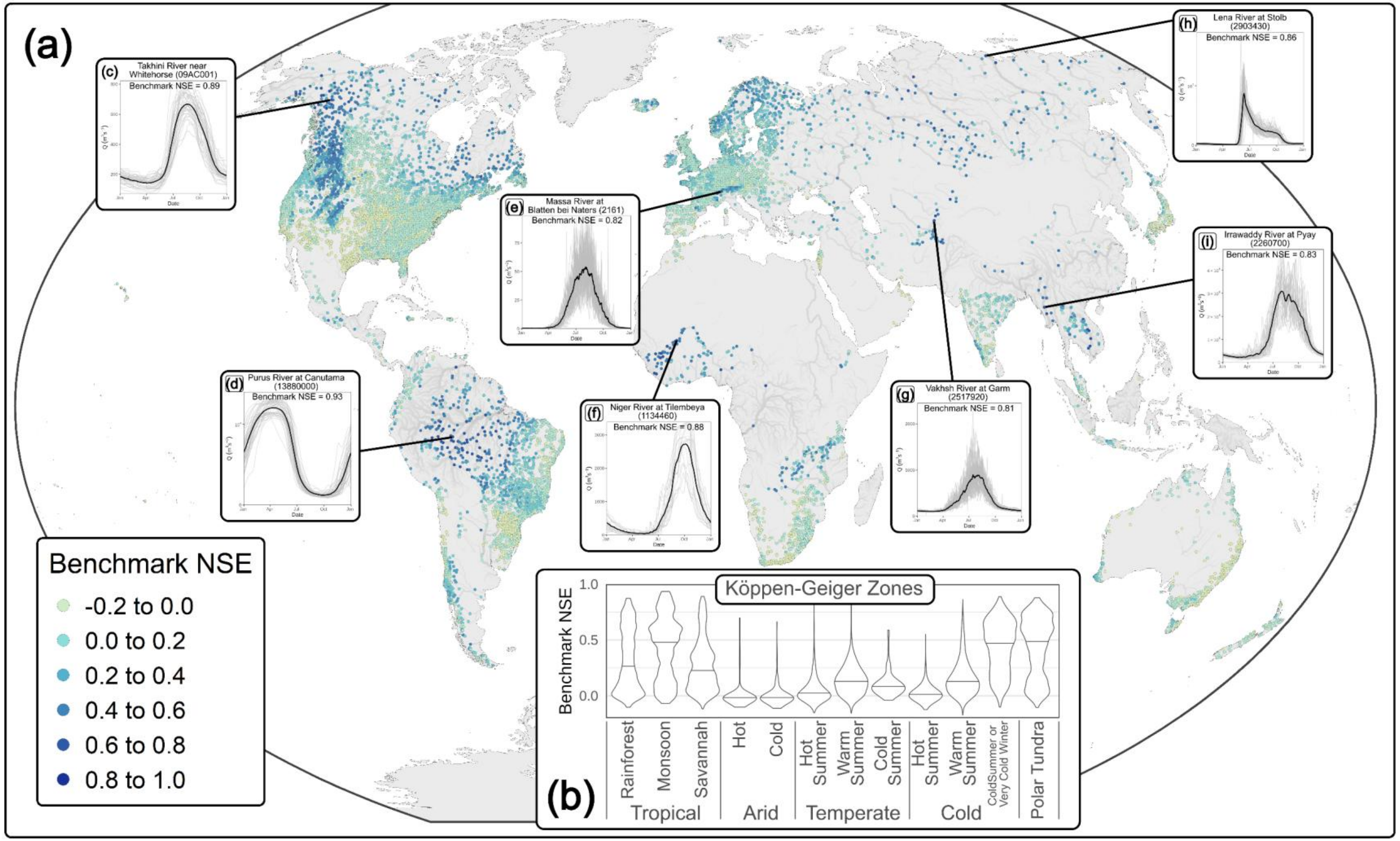
[1]University of Victoria, Canada
[2]University of Calgary, Canada
[3]University of Potsdam, Germany

What are ‘seasonal regimes’: where is the climatological benchmark NSE high?

Tropical monsoon

Alpine

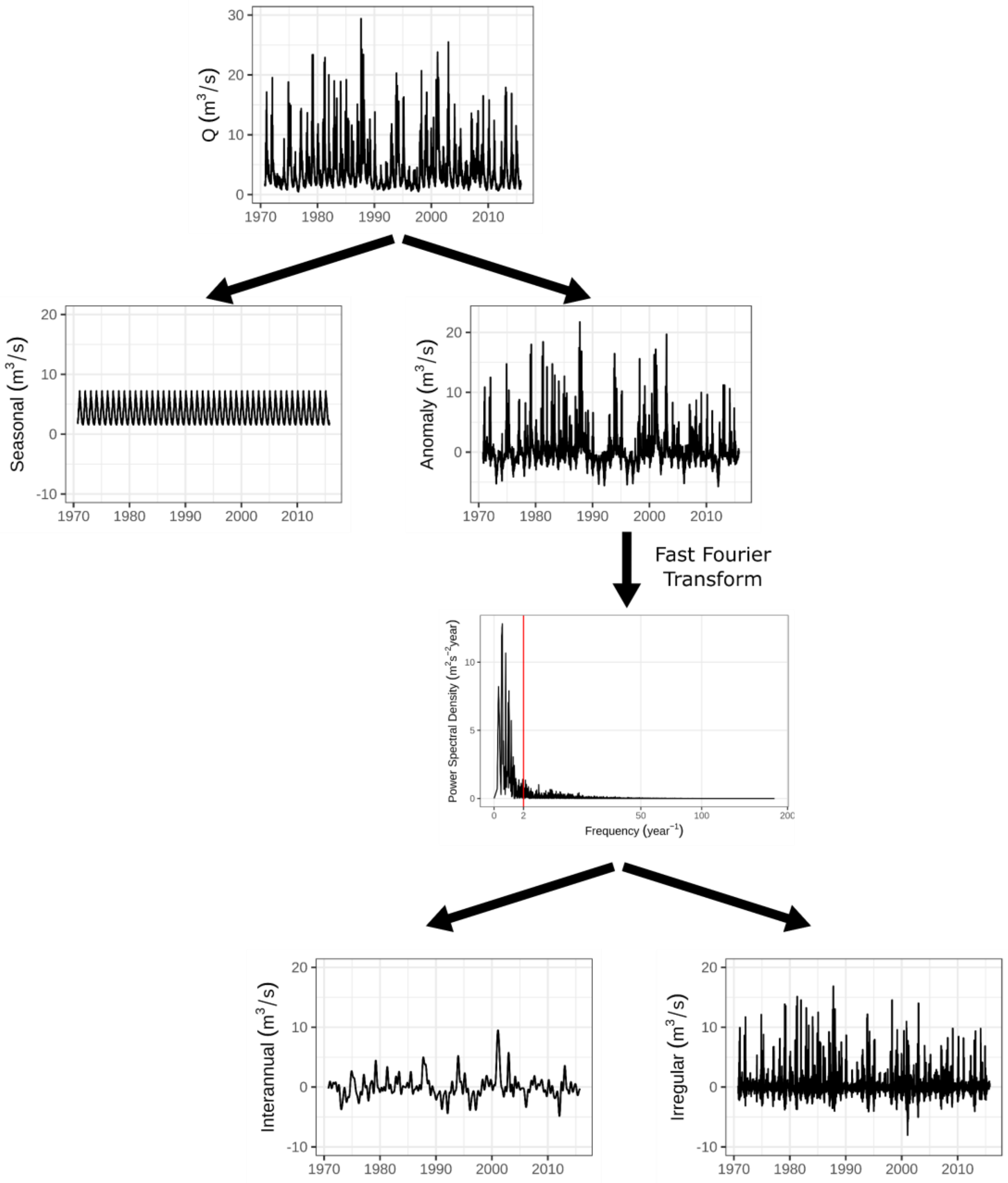
Polar



Streamflow time series decomposition

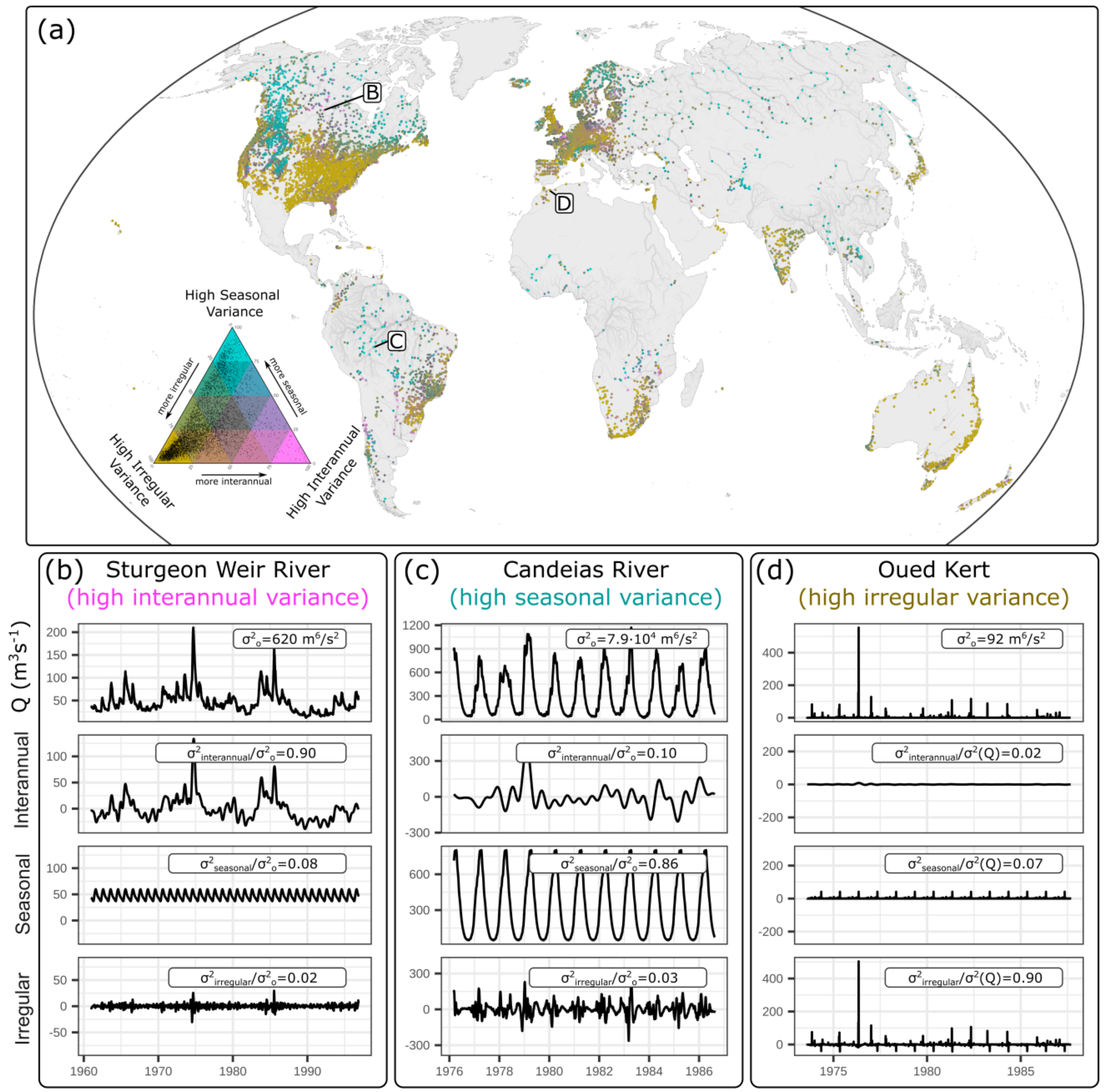
- Remove seasonal cycle and divide the remainder into high and low Fourier frequencies
- Decomposition is orthogonal:

$$\sigma^2(Q) = \sigma^2(Seasonal) + \sigma^2(Interannual) + \sigma^2(Irregular)$$



Seasonal, Interannual, and Irregular variance fractions

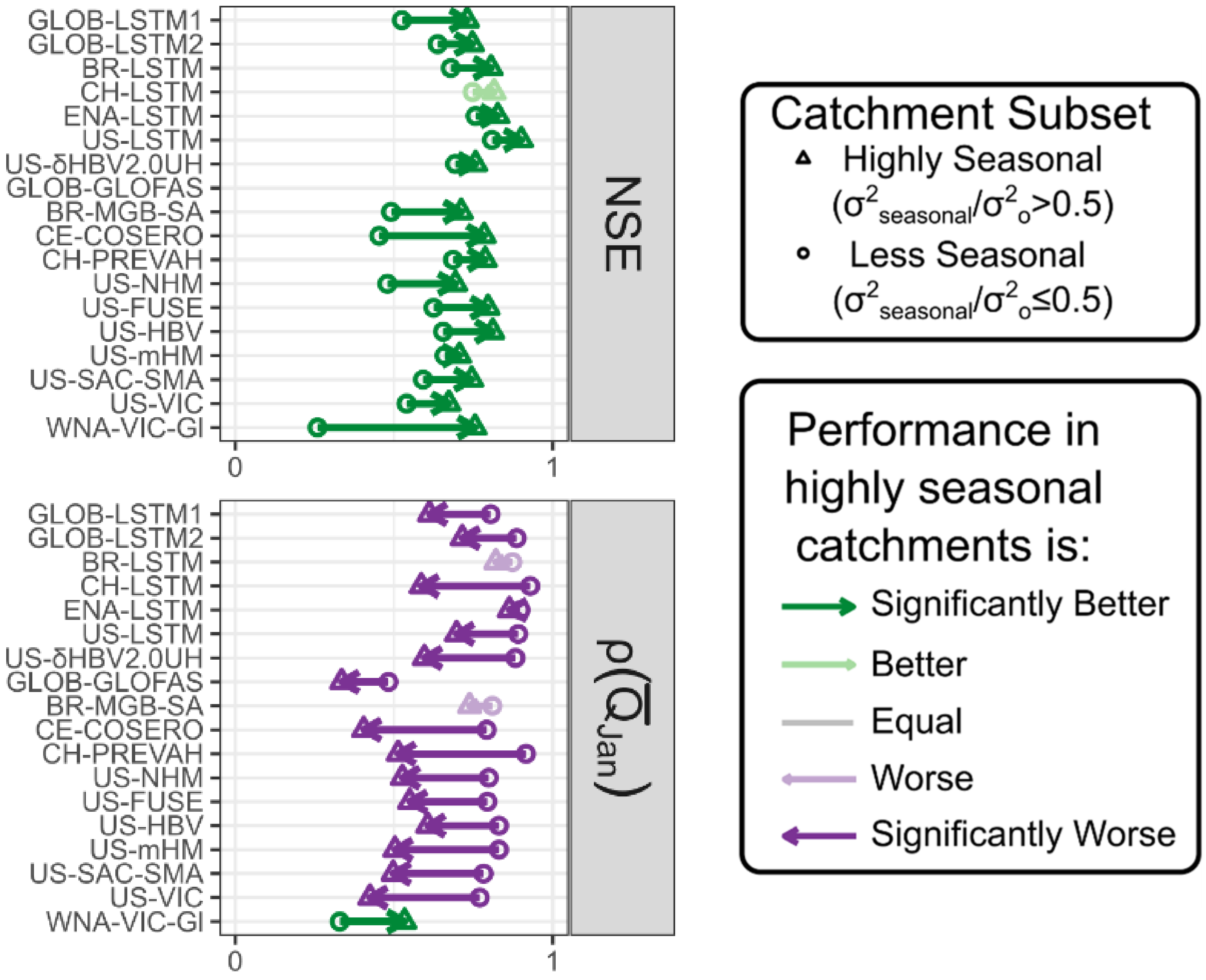
- Most catchments are irregular
- Highly seasonal catchments are found in tropical, alpine, or polar climates



High NSEs in seasonal catchments hide poor simulations of interannual variability

Across 18 models:

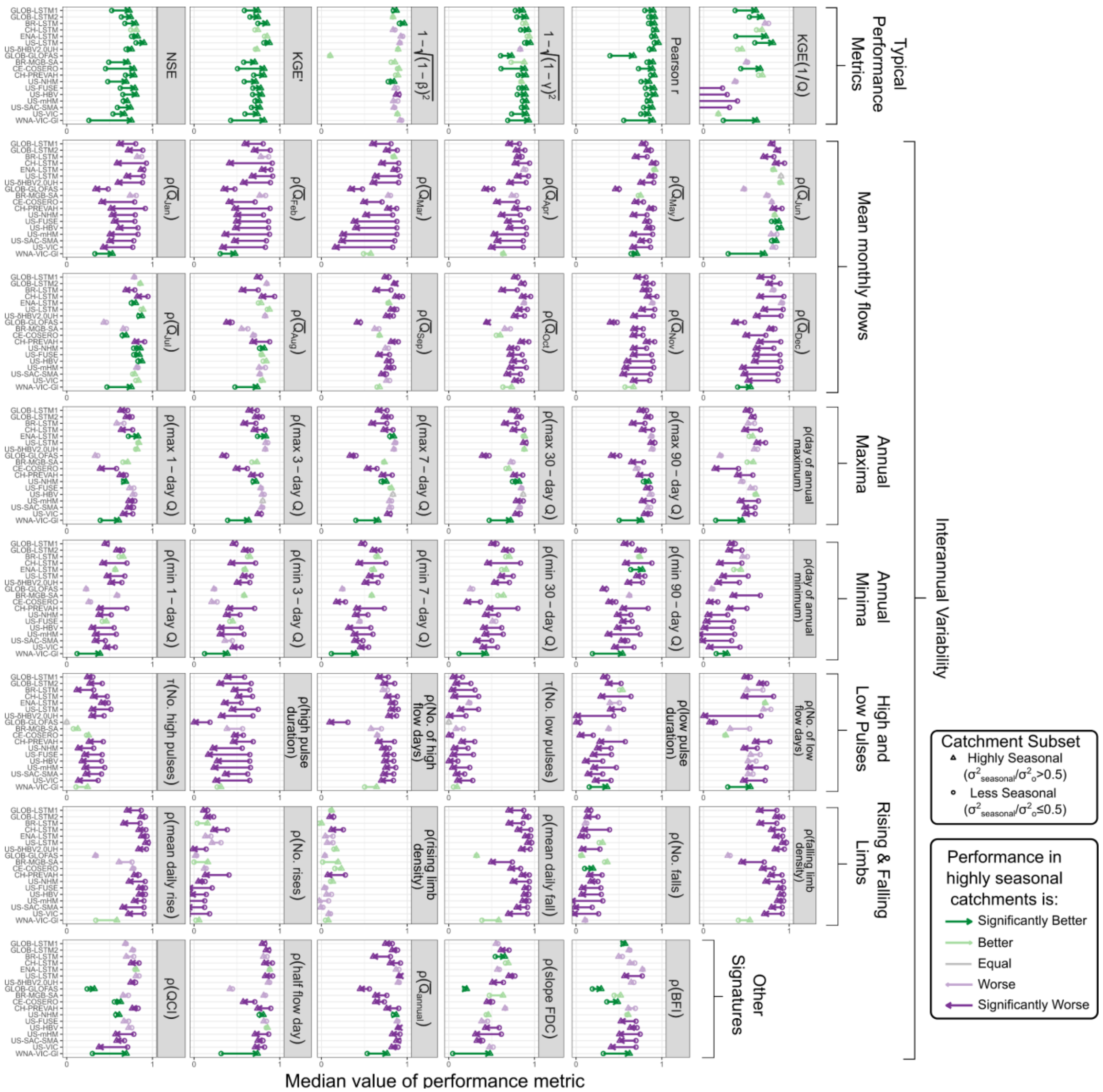
- NSEs are **always higher** in seasonal catchments
- Correlations with January flows are almost always worse



High NSEs in seasonal catchments hide poor simulations of interannual variability

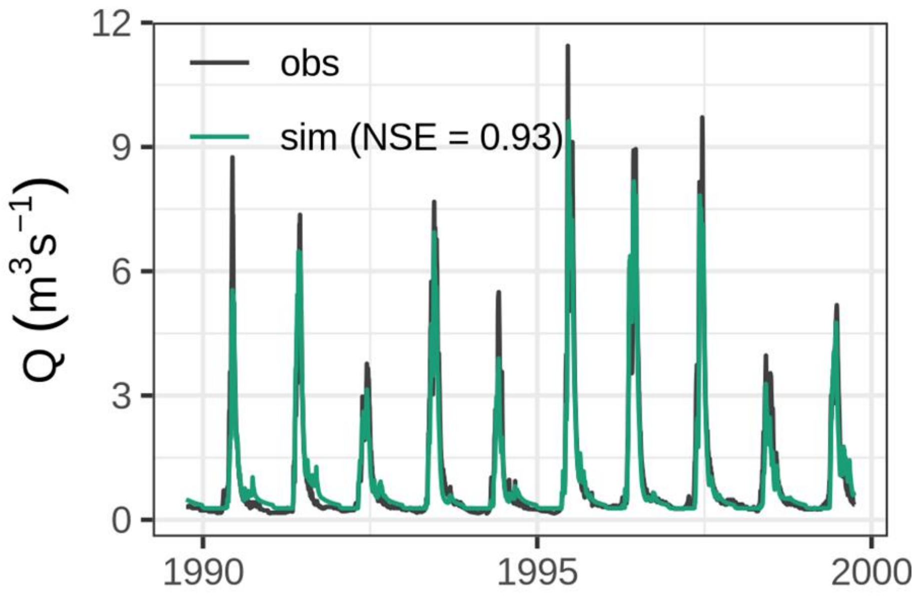
Across 18 models:

- Typical performance metrics are **mostly better** in seasonal catchments
- Simulation of year-to-year variation in hydrologic signatures is **usually worse**



Why: What explains this behaviour?

- **Two obvious reasons:**
- The NSE prioritizes high flows, so low-flow processes are poorly represented
- In seasonal regimes, the signal-to-noise ratio of the interannual variance is low, making model calibration and evaluation more difficult
- **But is that all?**
- Model optimization algorithms can bias training away from seasonal catchments
- Model structures often do not include the causes of interannual variance in seasonal catchments
- Data quantity and quality is often lowest in highly seasonal regimes (tropical, alpine, and polar climates)



Conclusion: Current generation of models struggle to simulate historical interannual variability in seasonal regimes

- Continued reliance on the NSE and KGE risks perpetuating this problem
- Raises concerns about using these models for forecasting and climate change projection

Model References

GLOB-LSTM1 and GLOB-GloFAS: Nearing, G. *et al.* Global prediction of extreme floods in ungauged watersheds. *Nature* **627**, 559–563 (2024).

GLOB-LSTM2: Yang, Y. *et al.* Global Daily Discharge Estimation Based on Grid Long Short-Term Memory (LSTM) Model and River Routing. *Water Resources Research* **61**, e2024WR039764 (2025).

BR-LSTM: Ruzzante, S. W., Knoben, W. J. M., Wagener, T., Gleeson, T. & Schnorbus, M. Technical note: High Nash–Sutcliffe Efficiencies conceal poor simulations of interannual variance in seasonal regimes. *Hydrology and Earth System Sciences* **30**, 2337–2355 (2026).

CH-LSTM and CH-PREVAH: Kraft, B. *et al.* CH-RUN: a deep-learning-based spatially contiguous runoff reconstruction for Switzerland. *Hydrology and Earth System Sciences* **29**, 1061–1082 (2025).

ENA-LSTM: Arsenault, R., Martel, J.-L., Brunet, F., Brissette, F. & Mai, J. Continuous streamflow prediction in ungauged basins: long short-term memory neural networks clearly outperform traditional hydrological models. *Hydrology and Earth System Sciences* **27**, 139–157 (2023).

US-LSTM: Kratzert, F., Gauch, M., Klotz, D. & Nearing, G. HESS Opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin. *Hydrology and Earth System Sciences* **28**, 4187–4201 (2024).

US-δHBV2.0UH: Song, Y. *et al.* High-Resolution National-Scale Water Modeling Is Enhanced by Multiscale Differentiable Physics-Informed Machine Learning. *Water Resources Research* **61**, e2024WR038928 (2025).

BR-MGB-SA: Siqueira, V. A. *et al.* Toward continental hydrologic–hydrodynamic modeling in South America. *Hydrology and Earth System Sciences* **22**, 4815–4842 (2018).

CE-COSERO: Klingler, C., Schulz, K. & Herrnegger, M. LamaH-CE: LArge-SaMple DAta for Hydrology and Environmental Sciences for Central Europe. *Earth System Science Data* **13**, 4529–4565 (2021).

US-NHM: Regan, R. S. *et al.* The U. S. Geological Survey National Hydrologic Model infrastructure: Rationale, description, and application of a watershed-scale model for the conterminous United States. *Environmental Modelling & Software* **111**, 192–203 (2019).

US-FUSE: Kratzert, F. CAMELS benchmark models. HydroShare https://doi.org/10.4211/hs.474ecc37e7db45baa425cdb4fc1b61e1 (2019).

US-HBV: Seibert, J., Vis, M. J. P., Lewis, E. & van Meerveld, H. j. Upper and lower benchmarks in hydrological modelling. *Hydrological Processes* **32**, 1120–1125 (2018).

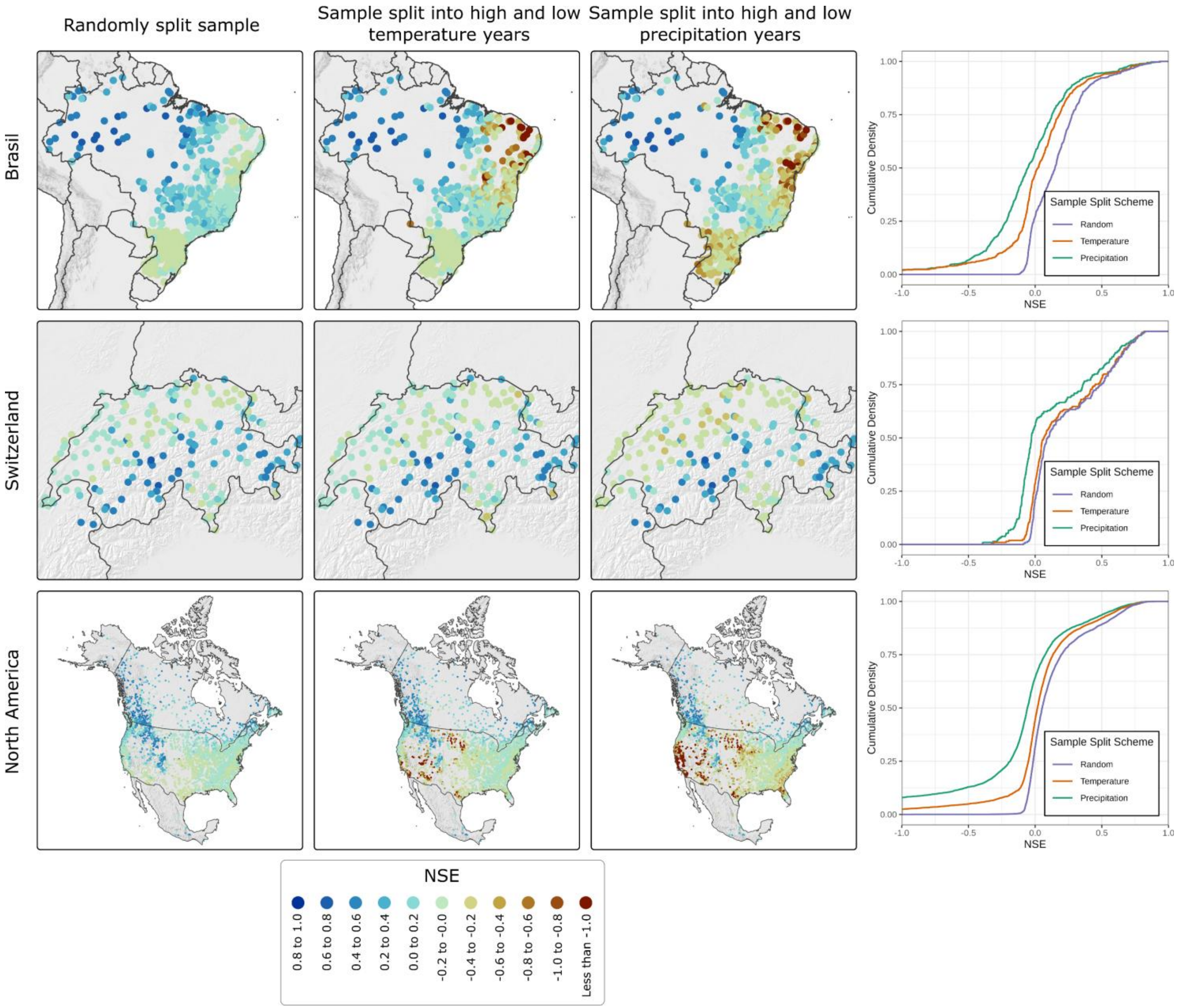
US-mHM: Mizukami, N. *et al.* On the choice of calibration metrics for “high-flow” estimation using hydrologic models. *Hydrology and Earth System Sciences* **23**, 2601–2614 (2019).

US-SAC-SMA: and US-VIC: Newman, A. J. *et al.* Benchmarking of a Physically Based Hydrologic Model. https://doi.org/10.1175/JHM-D-16-0284.1 (2017) doi:10.1175/JHM-D-16-0284.1.

WNA-VIC-GL: Schnorbus, M. *VIC Glacier (VIC-GL) - Description of VIC Model Changes and Upgrades, VIC Generation 2 Deployment Report Volume 1*. 40 https://dspace.library.uvic.ca/server/api/core/bitstreams/850d5363-e326-4b13-bddc-f5cd6227c928/content (2018).

Does using a differential split sample make the NSE a reliable performance indicator?

No: The climatological benchmark NSE remains high under differential splitting in seasonal regimes

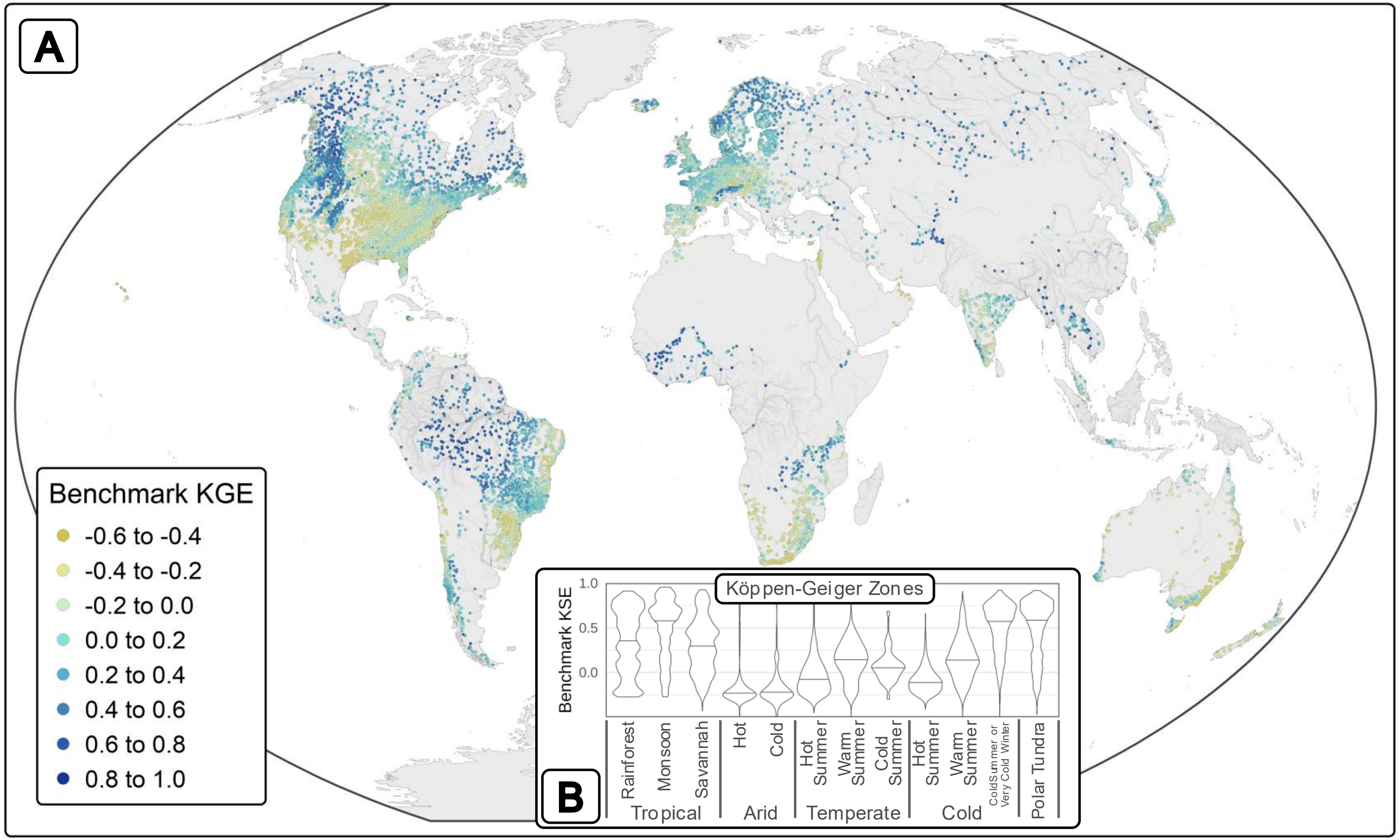


Where is the climatological benchmark **KGE** high?

Tropical monsoon, alpine, and polar regions – same as NSE

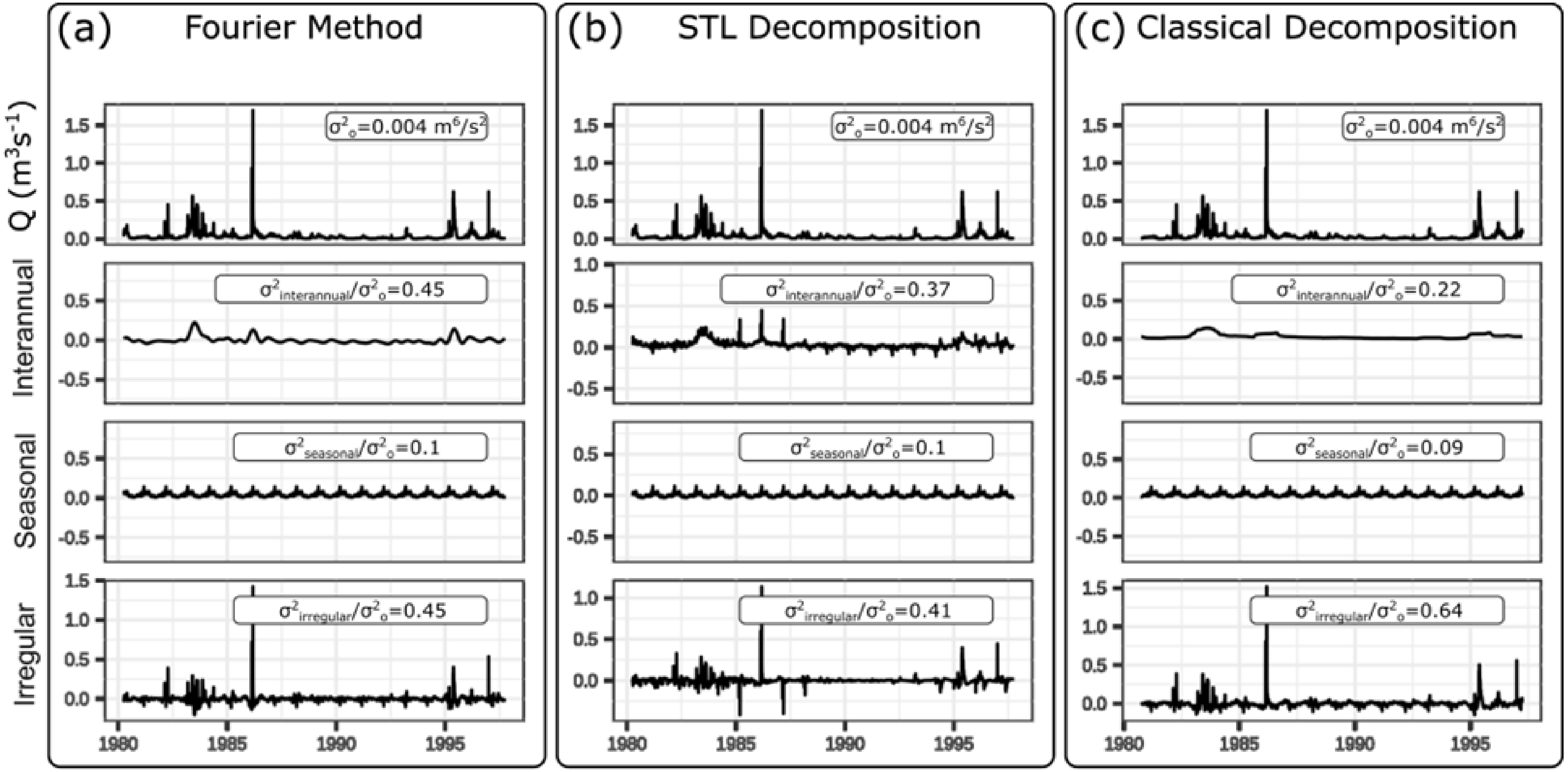
In fact, the benchmark KGE is related to the benchmark NSE by:

$$KGE_{cb} = 1 - \sqrt{2} + \sqrt{2 \times NSE_{cb}}$$

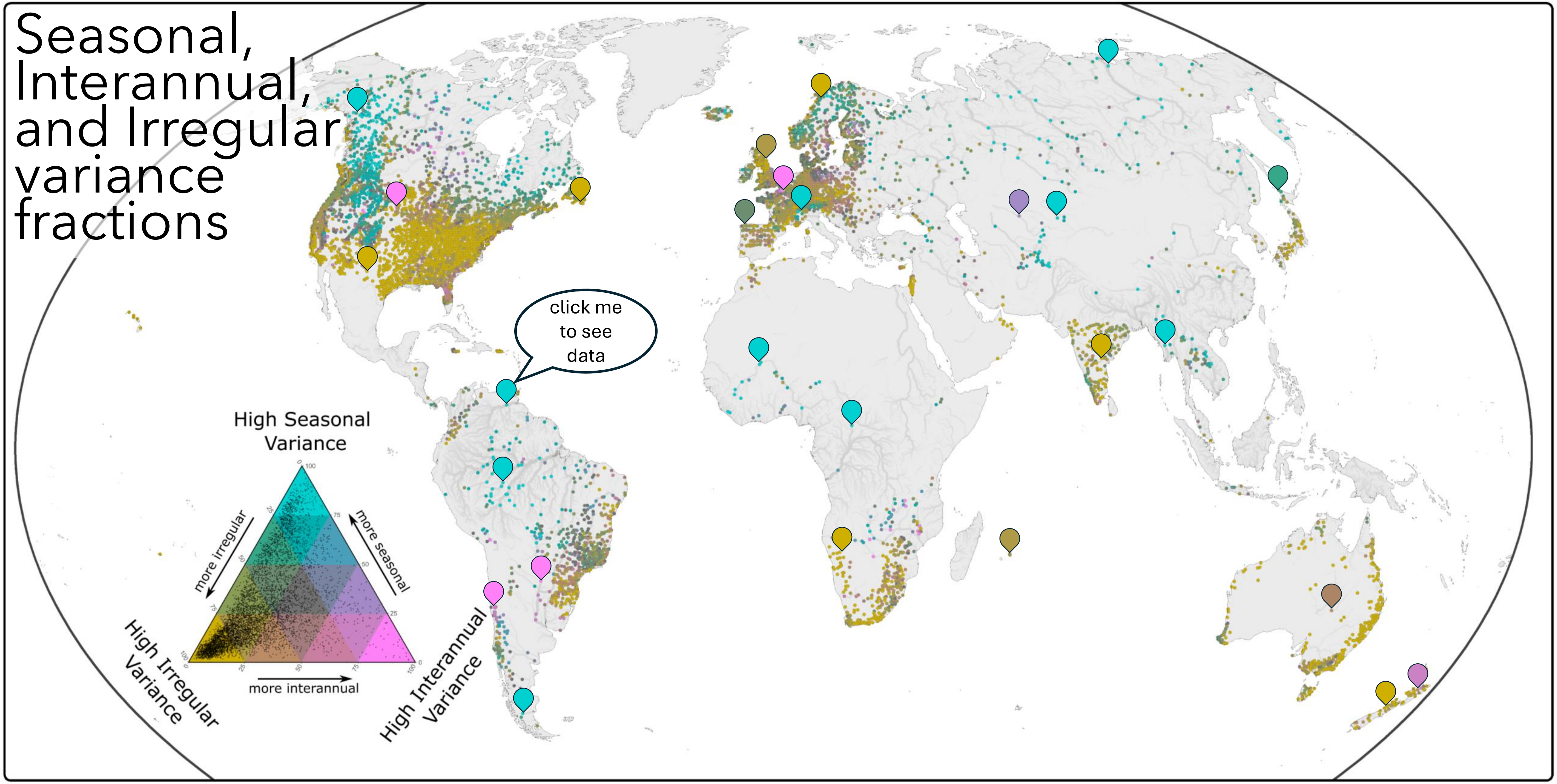


What about other decomposition methods?

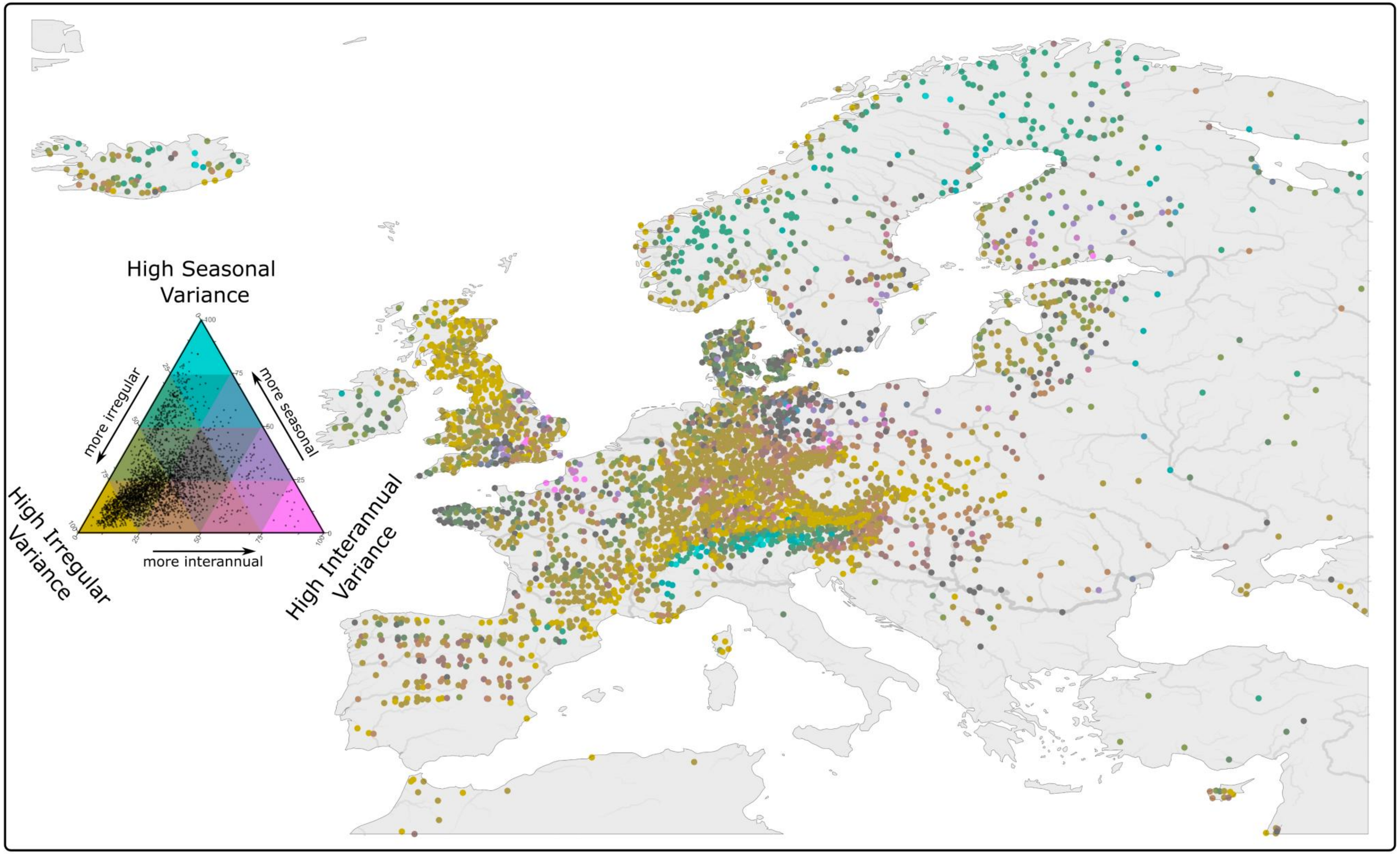
- Classical and STL decomposition give similar results
- Classical decomposition tends to underestimate interannual variance and STL tends to overestimate it.
- Classical and STL variance fractions do not sum to 1



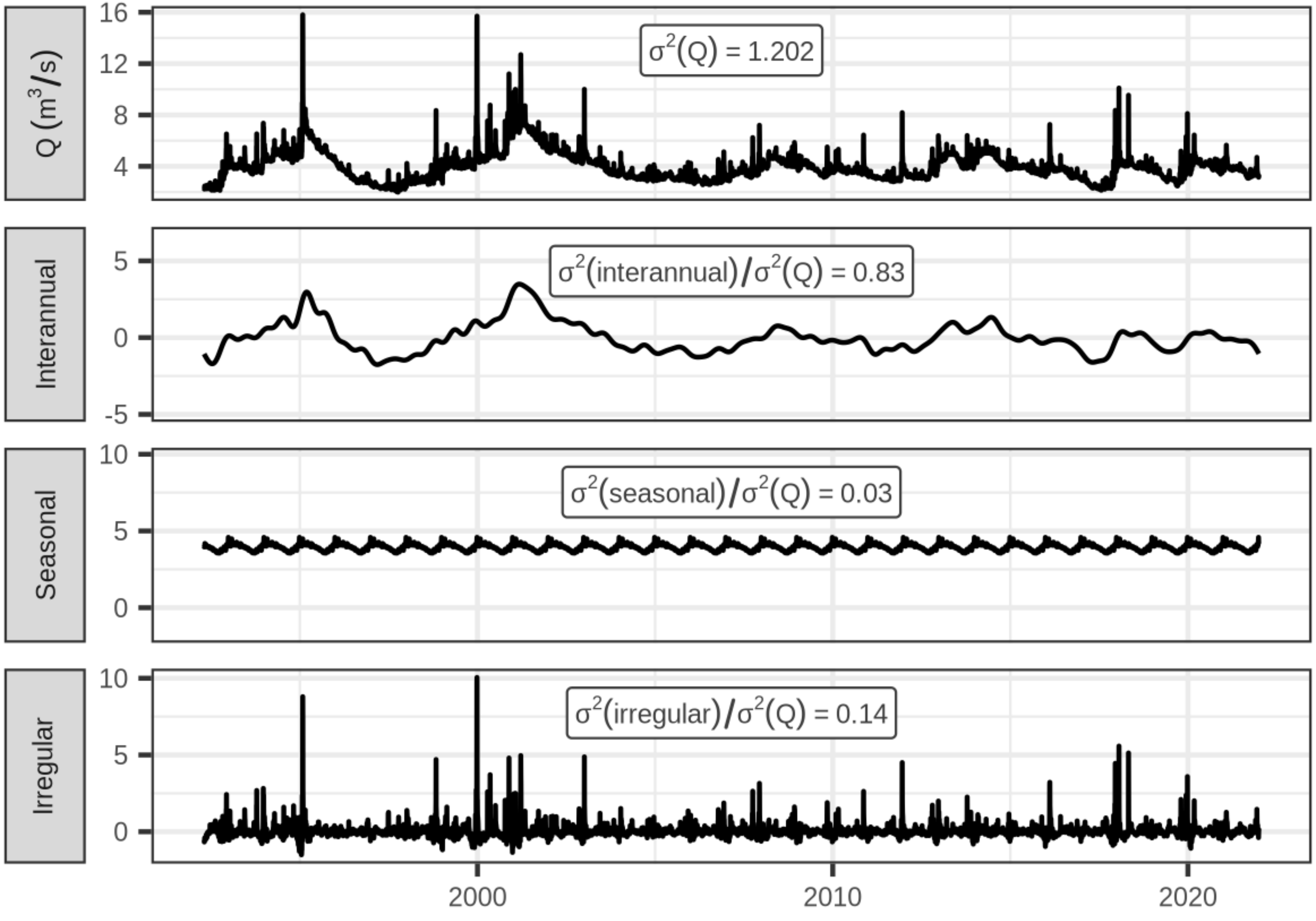
Seasonal, Interannual, and Irregular variance fractions



Seasonal, Interannual, and Irregular variance fractions - Europe

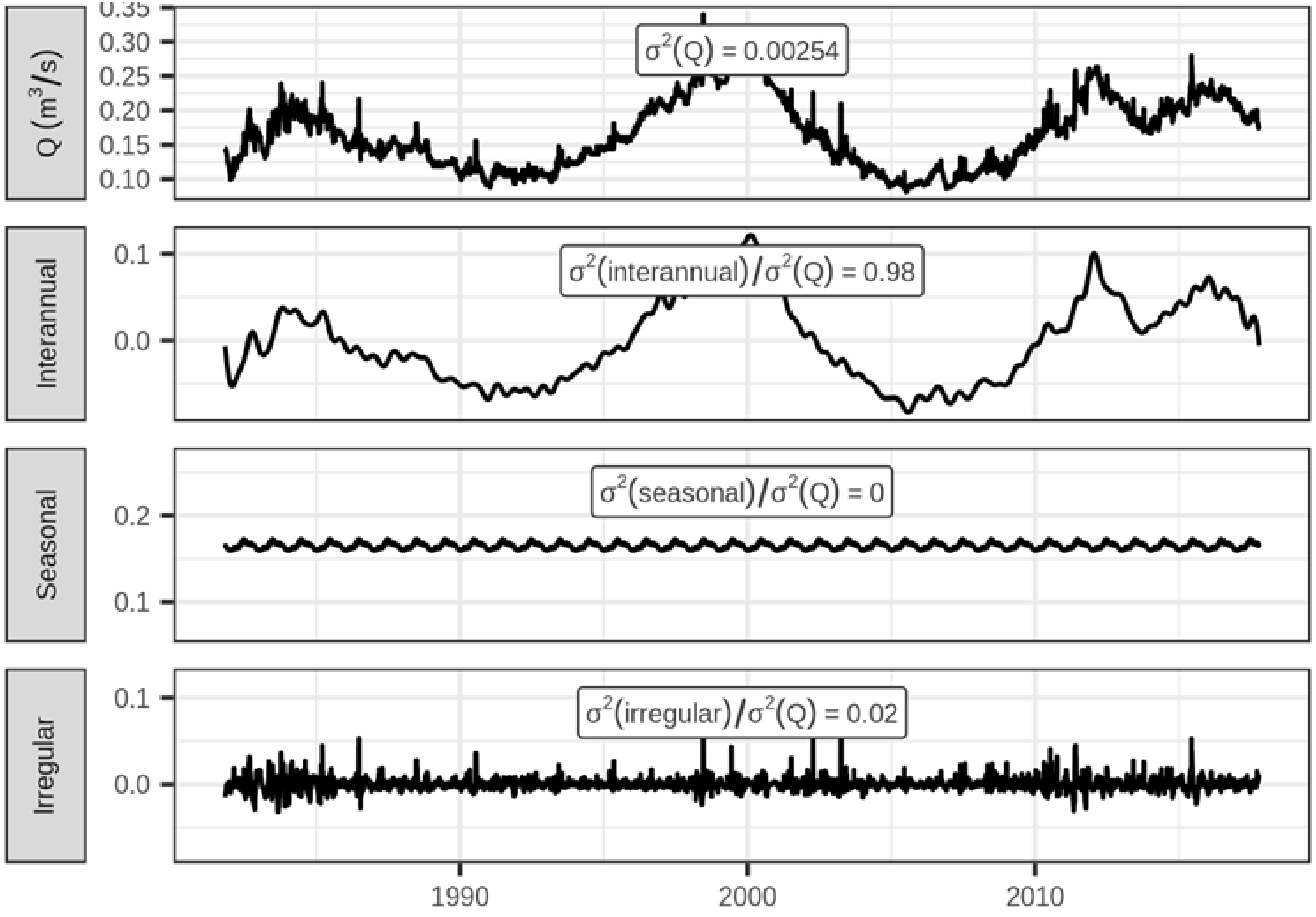


La Durdent à Vittefleur (Sandre G600061010)



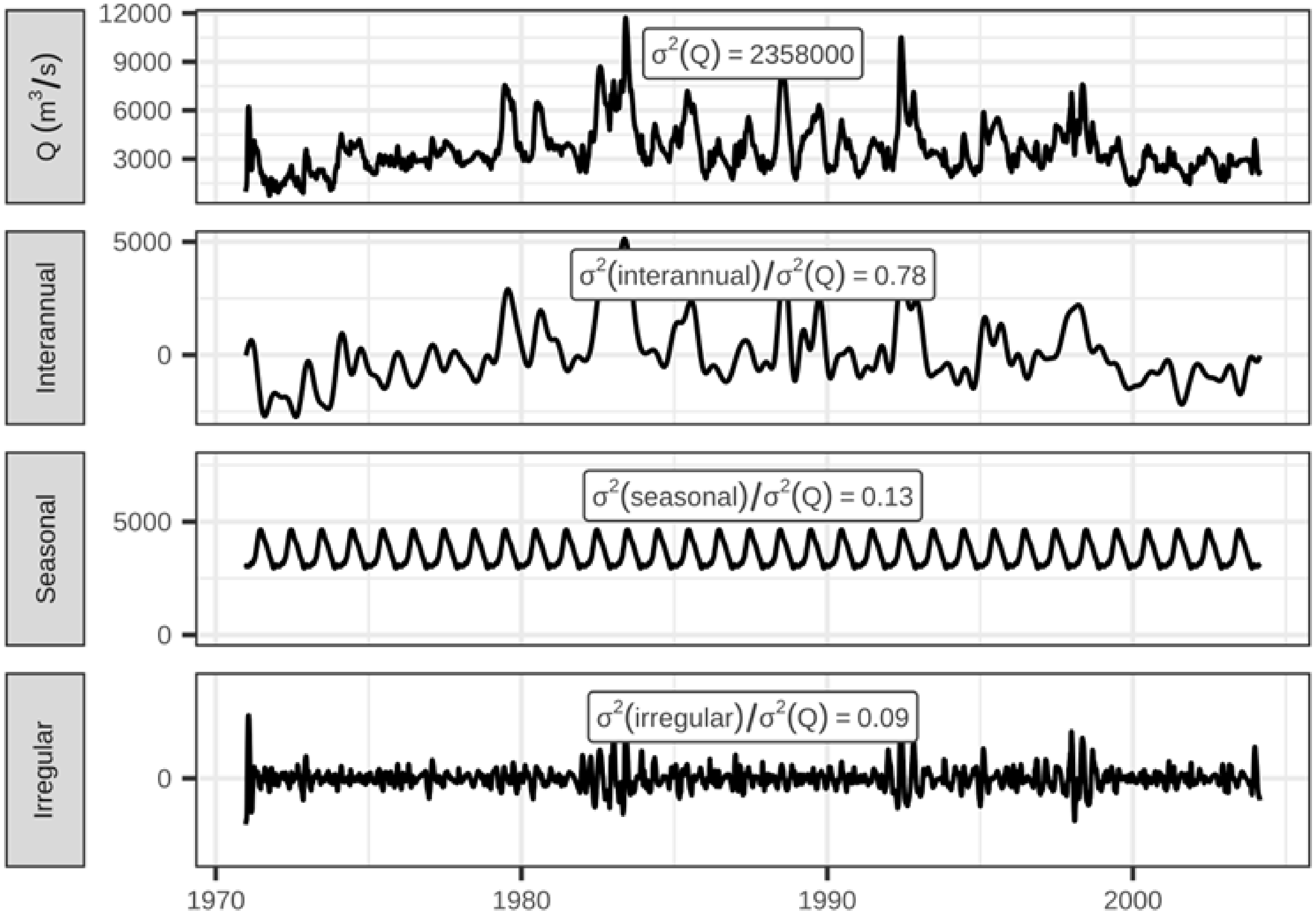
back to map:

Rhoads Fork Near Rochford, SD (USGS 06408700)



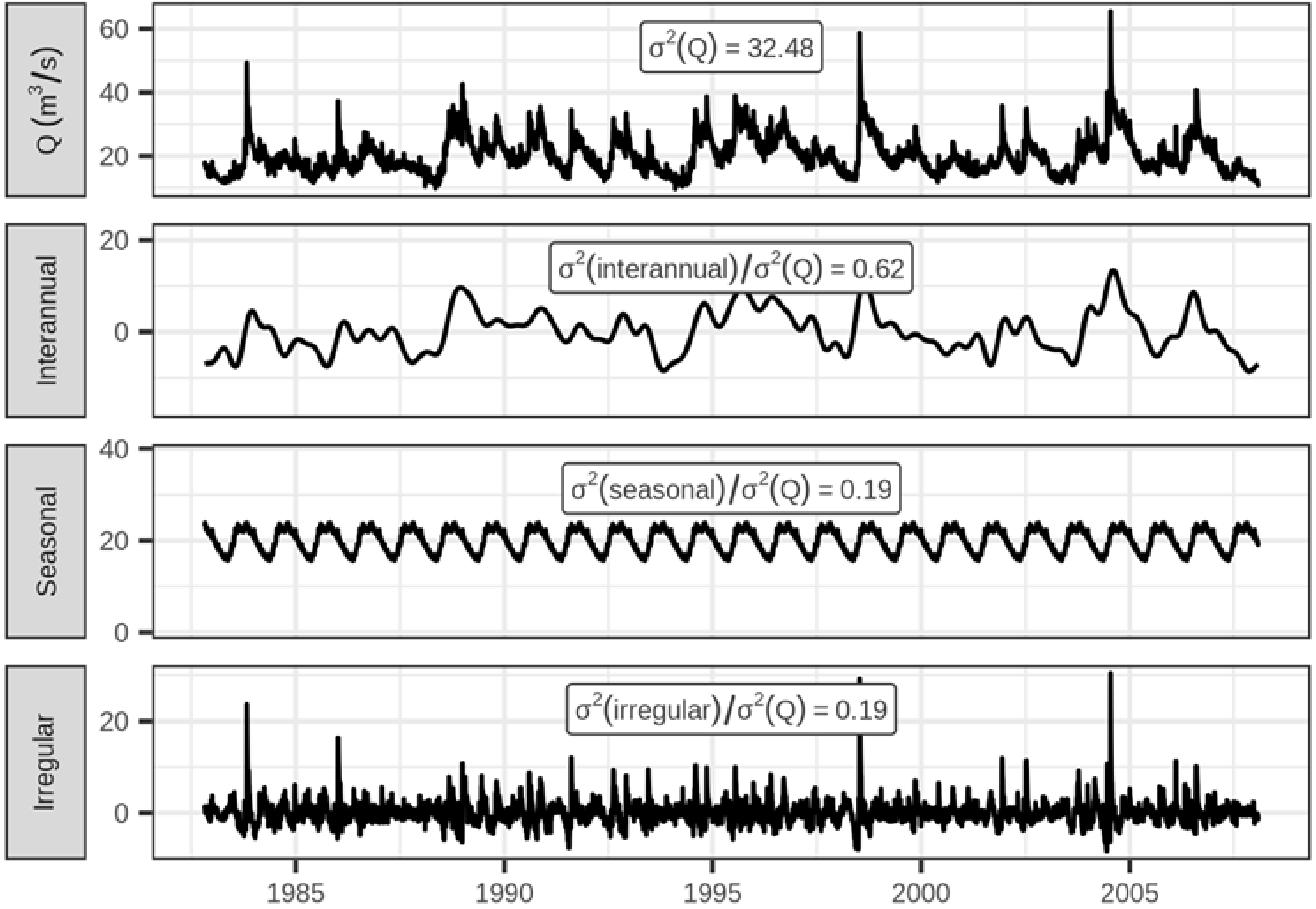
back to map:

Ysyry Paraguái (Paraguay River) at Asunción, Paraguay, (GRDC 3368100)

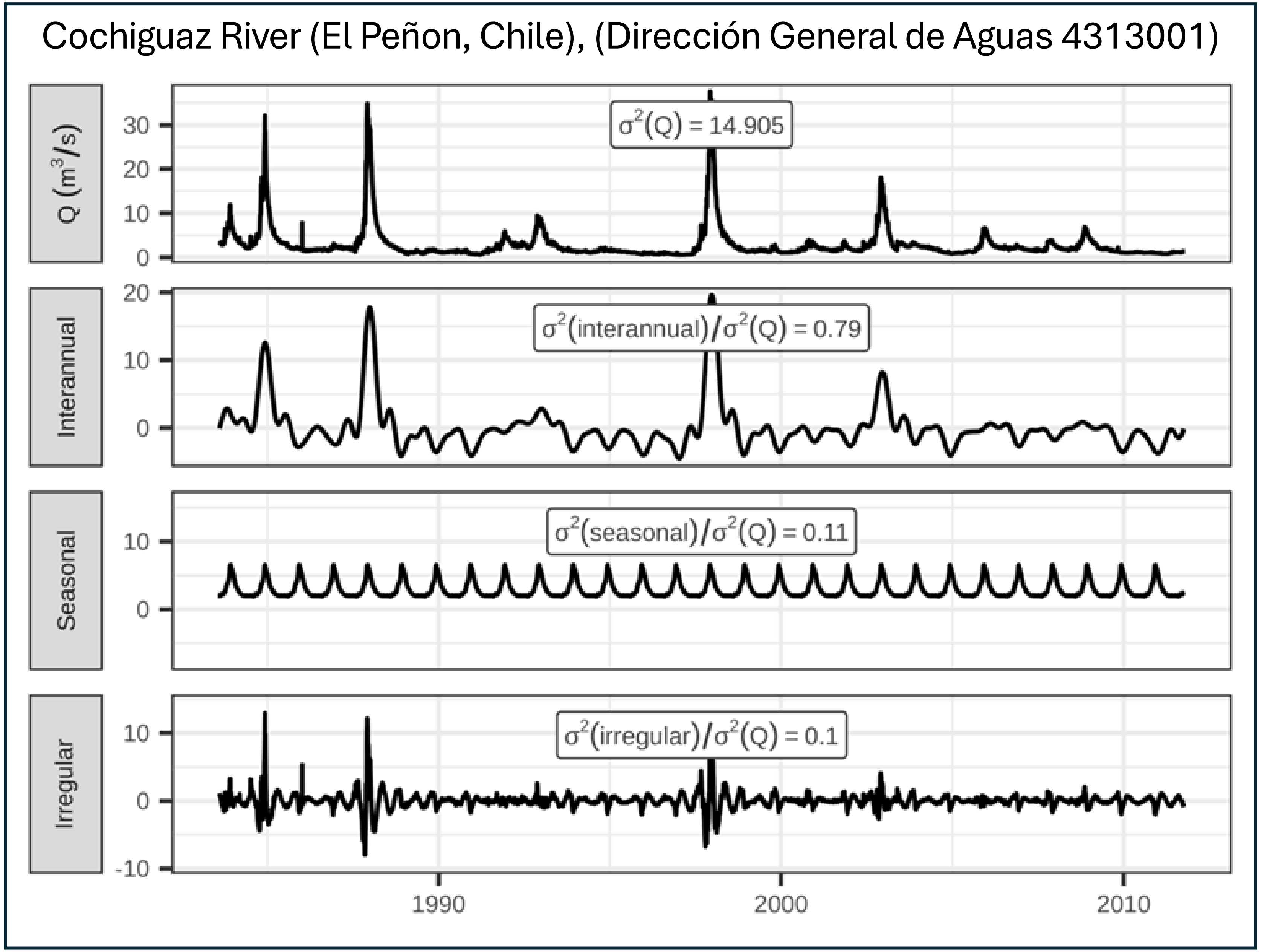


back to map:

Rangitaiki River at Murupara, Aotearoa (New Zealand), (GRDC 5863120)

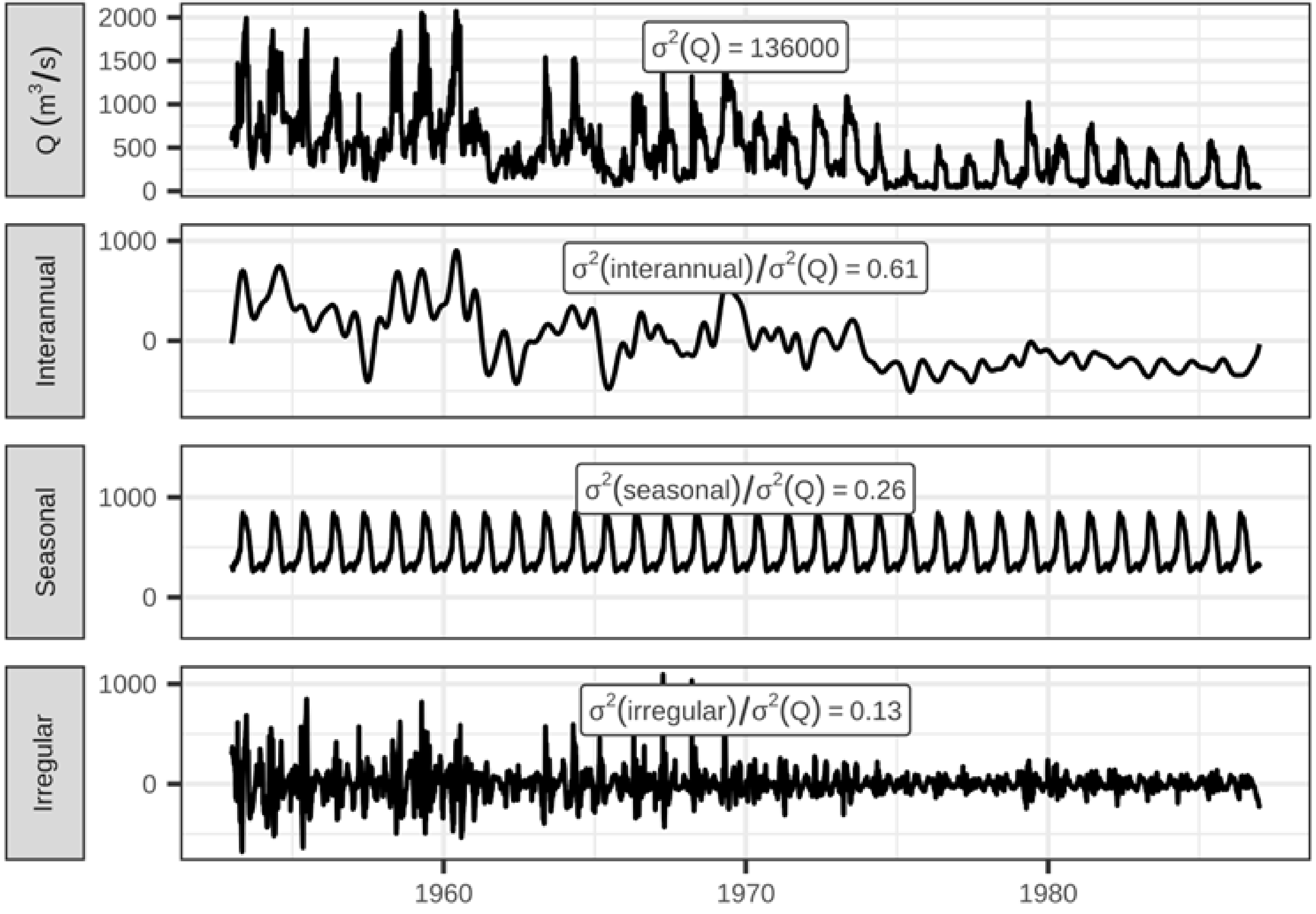


back to map:


Cochiguaz River (El Peñon, Chile), (Dirección General de Aguas 4313001)
Q (m³/s)
Interannual
Seasonal
Irregular
σ²(Q) = 14.905
σ²(interannual)/σ²(Q) = 0.79
σ²(seasonal)/σ²(Q) = 0.11
σ²(irregular)/σ²(Q) = 0.1
1990
2000
2010

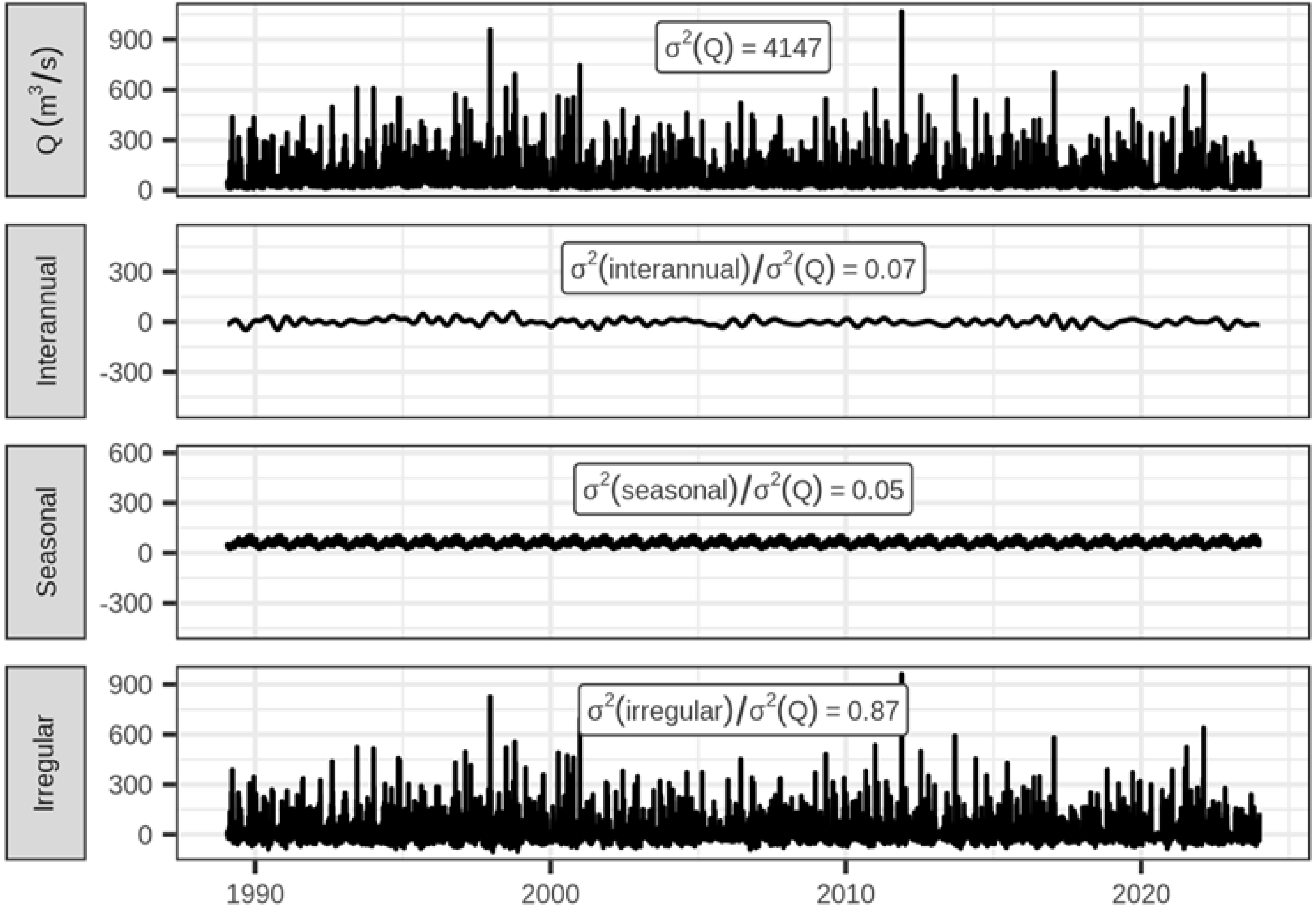
back to map:

Syr Darya (Tyumen-Aryk, Kazakhstan), (GRDC 2316200)



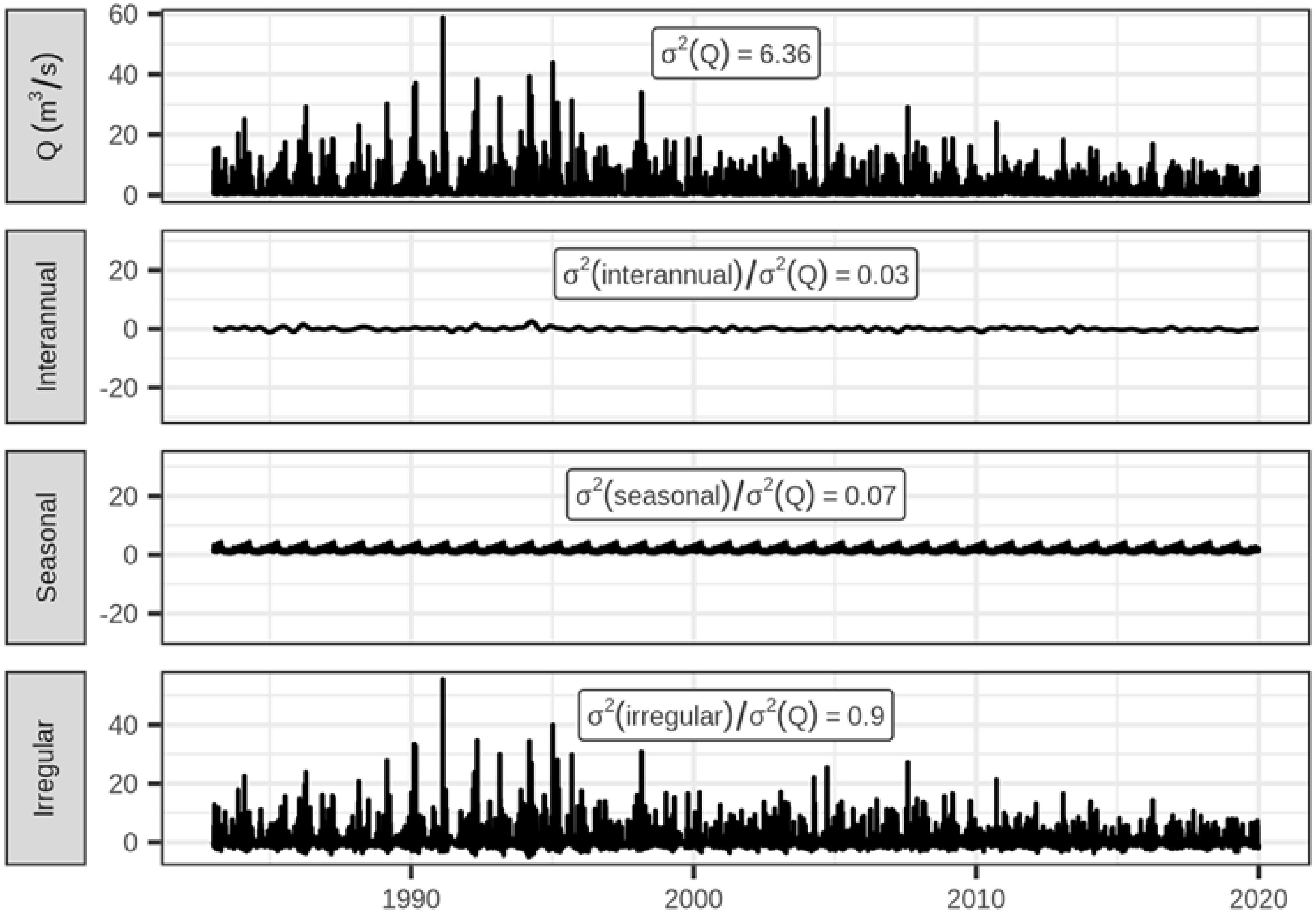
back to map:

Māwheranui River (New Zealand), (GRDC 5867710)



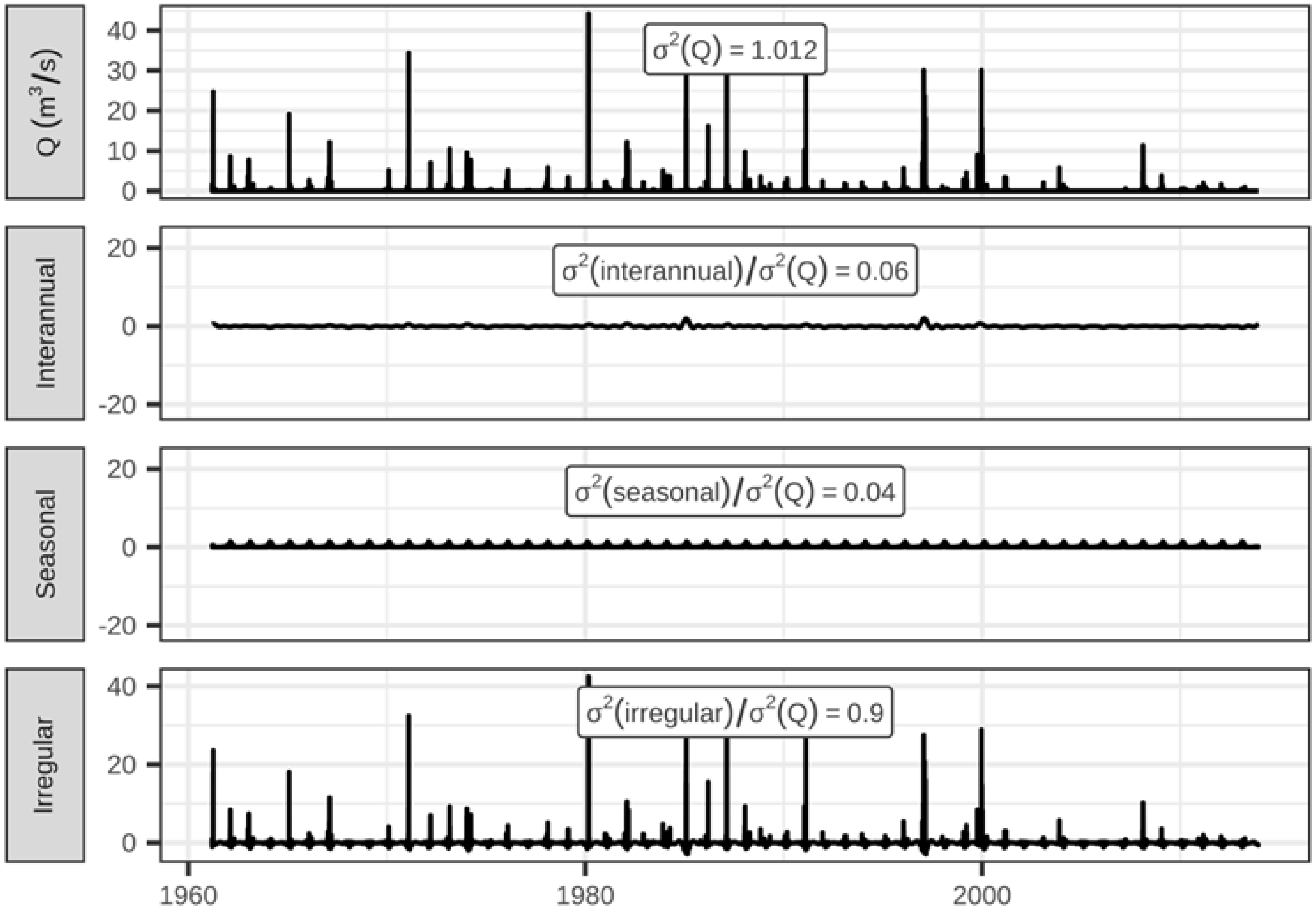
back to map:

Little Barachois River (Newfoundland, Canada), (Water Survey of Canada 02ZK003)



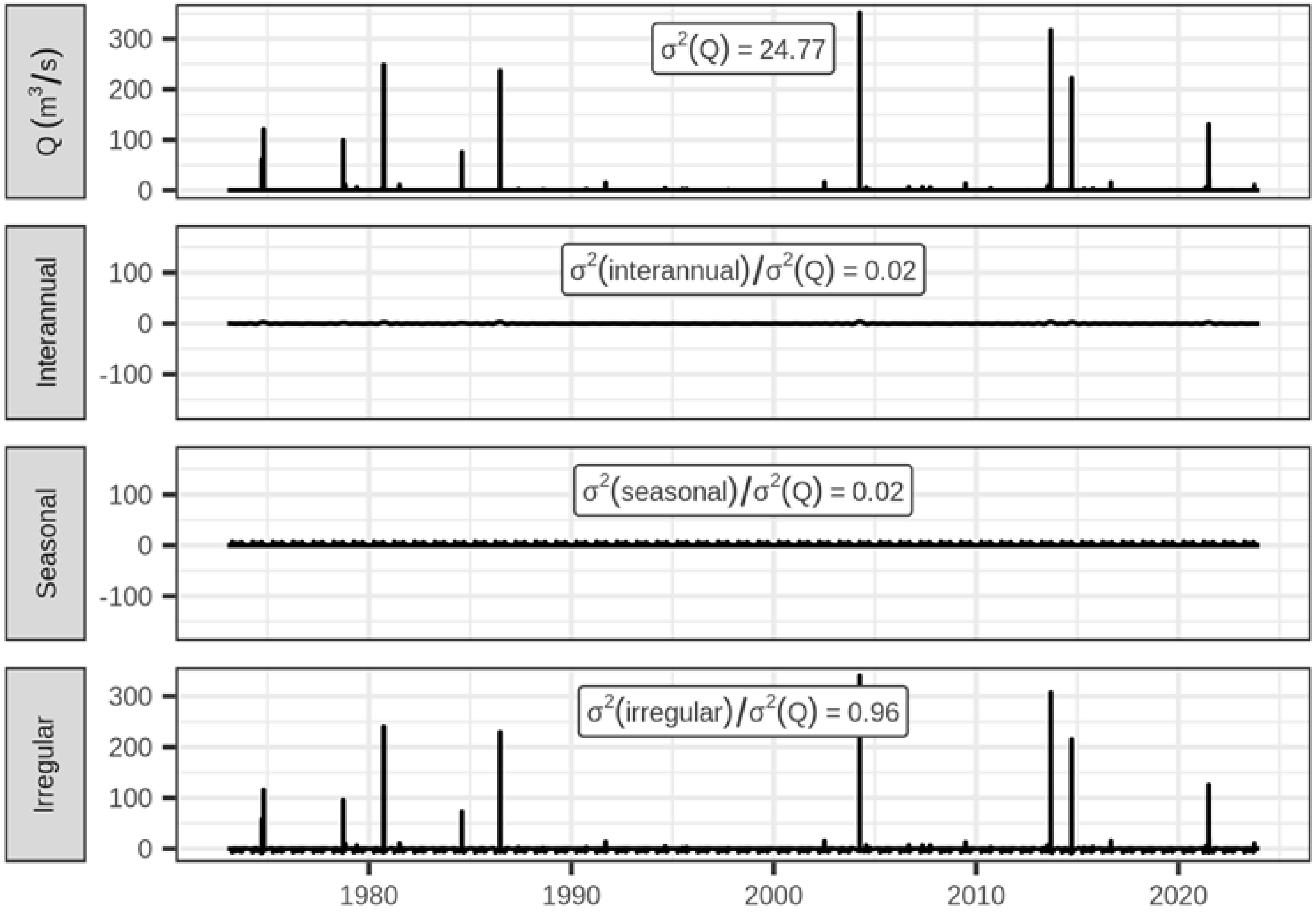
back to map:

Ugab River (Namibia), (GRDC 1258202)



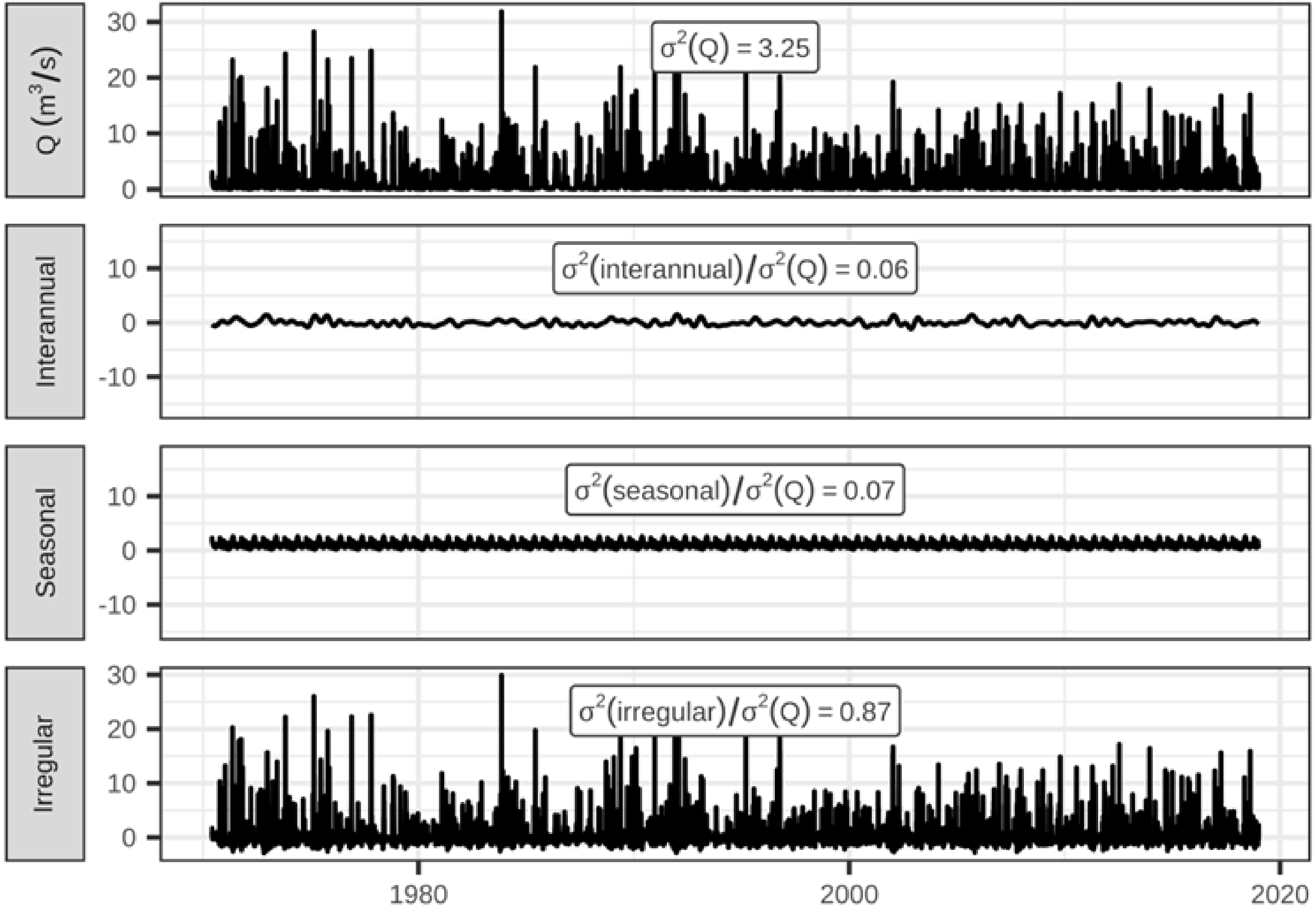
back to map:

Dark Canyon at Carlsbad (New Mexico, United States), (USGS 08405150)



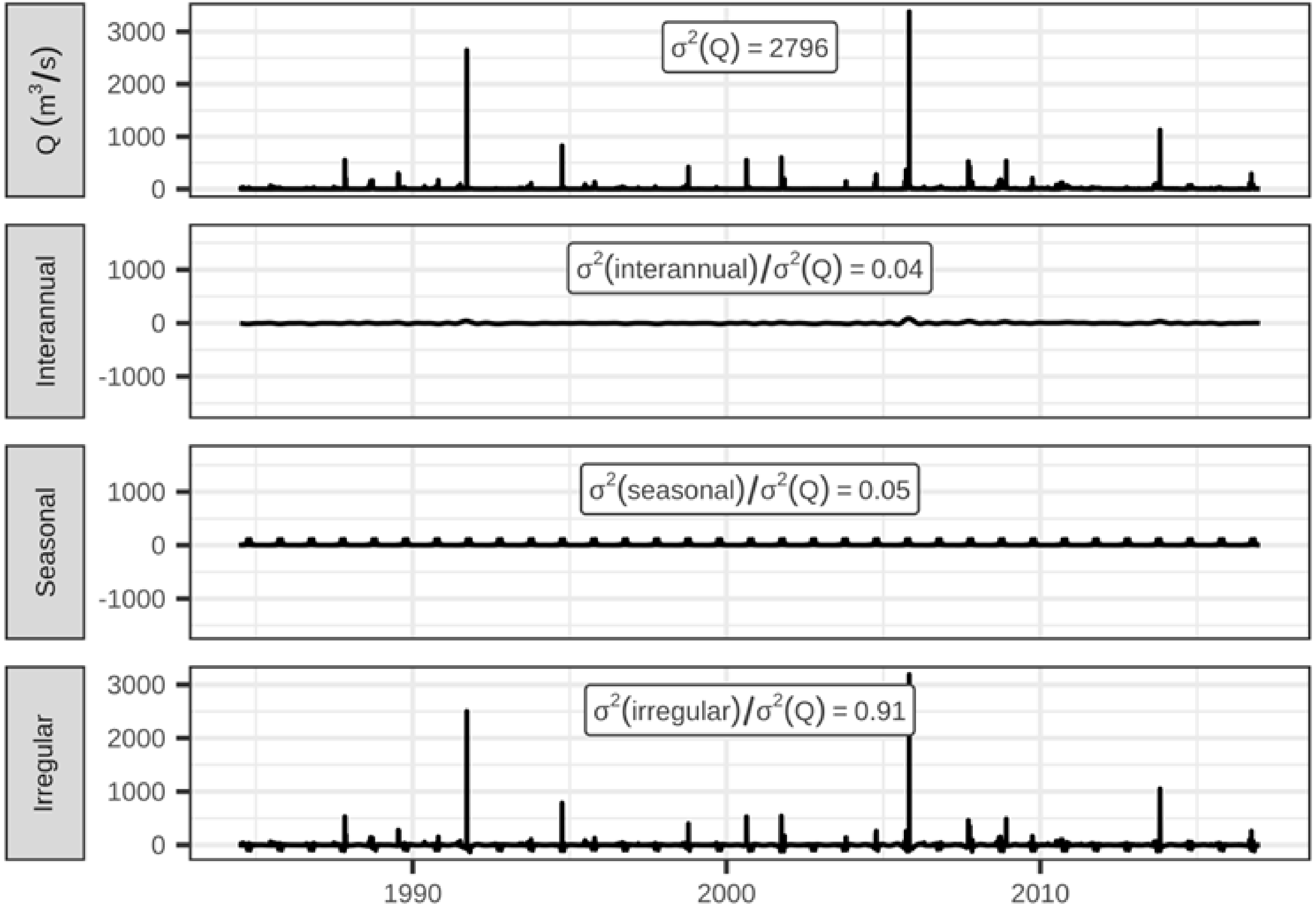
back to map:

Gieddejohka River (Leirpoldvatn, Norway), (GRDC 6731750)



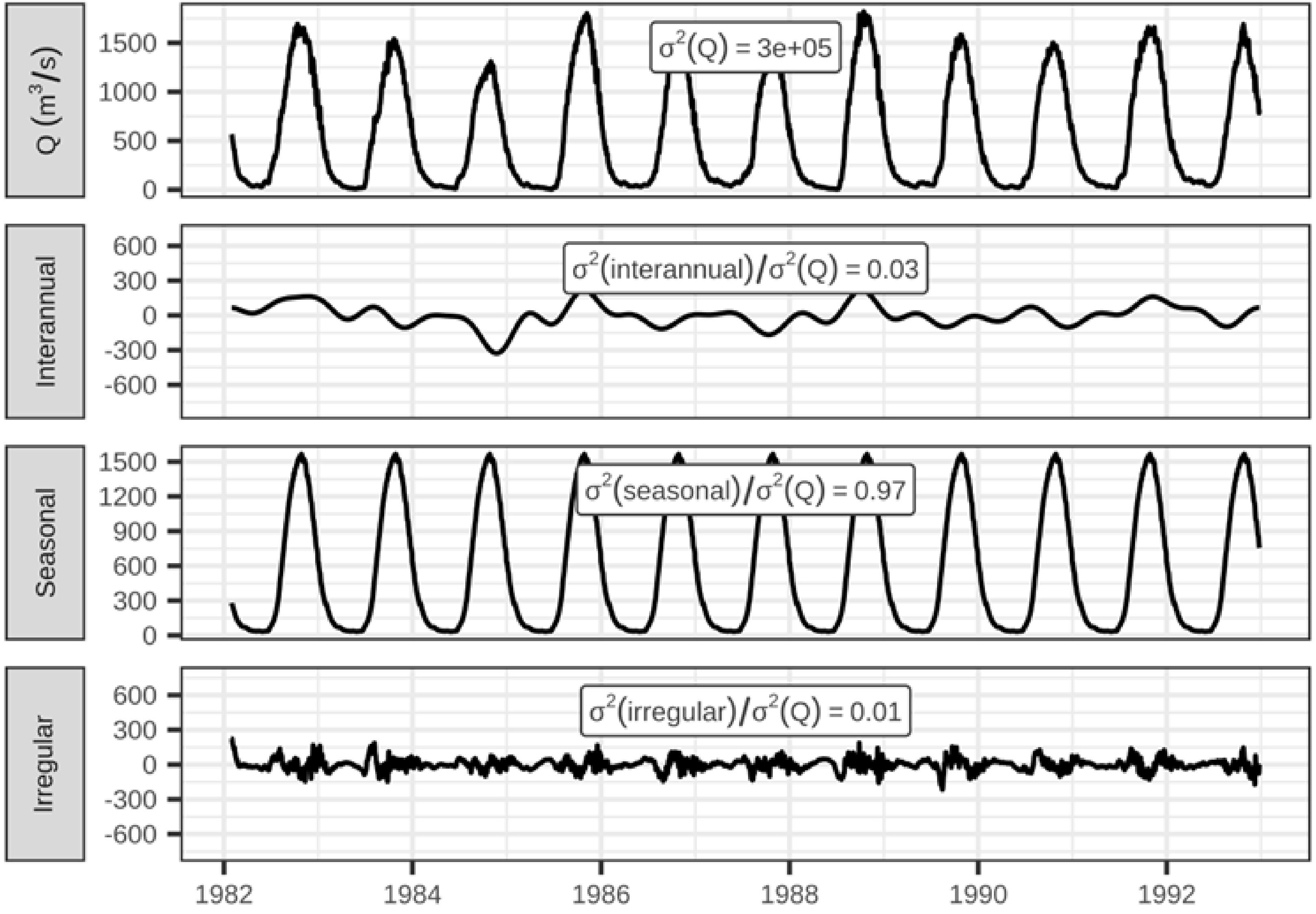
back to map:

Haliya River (Telangana, India), (Camels-IND 04012)



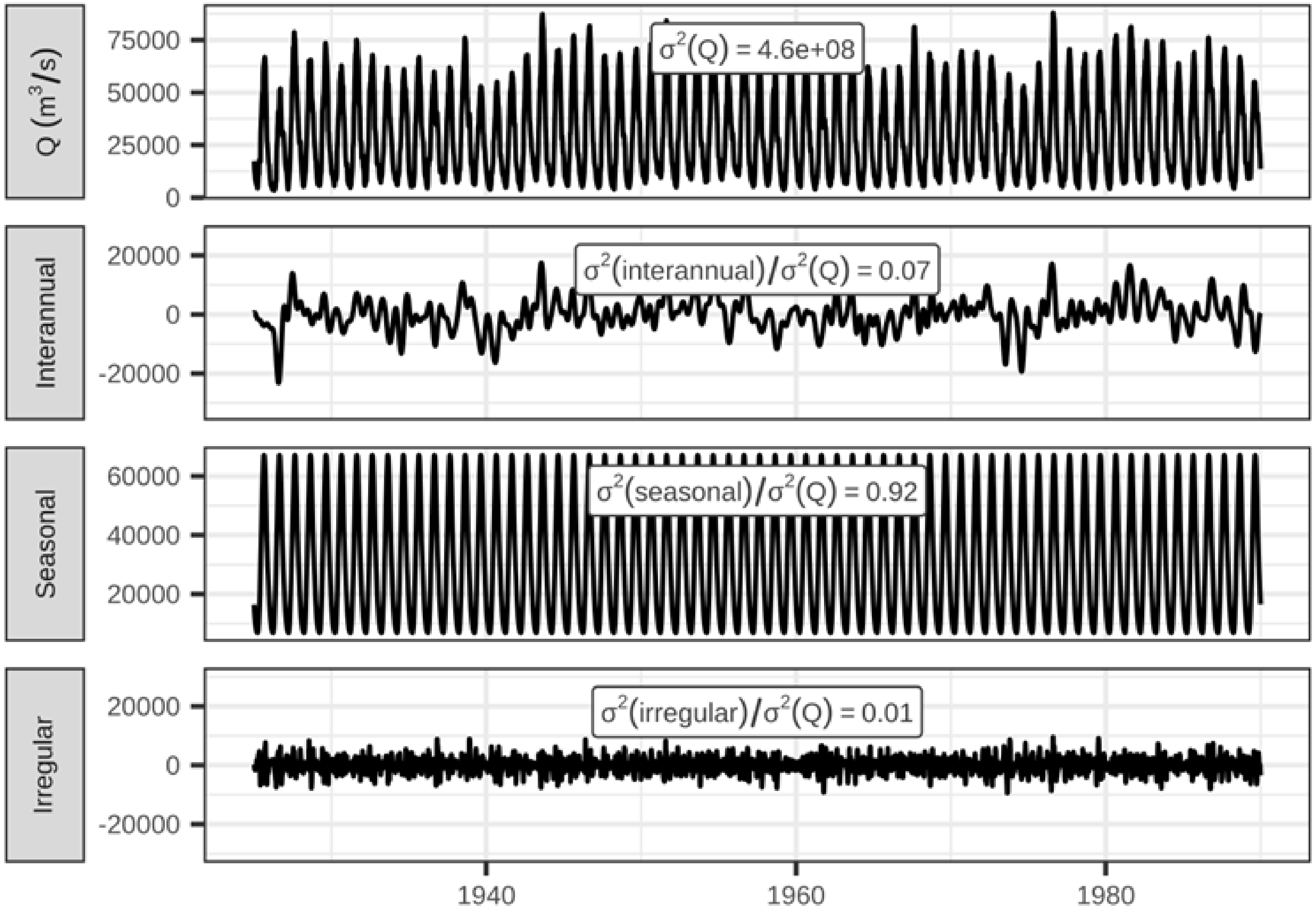
back to map:

Niger River at Dire (Mali), (GRDC 1134700)

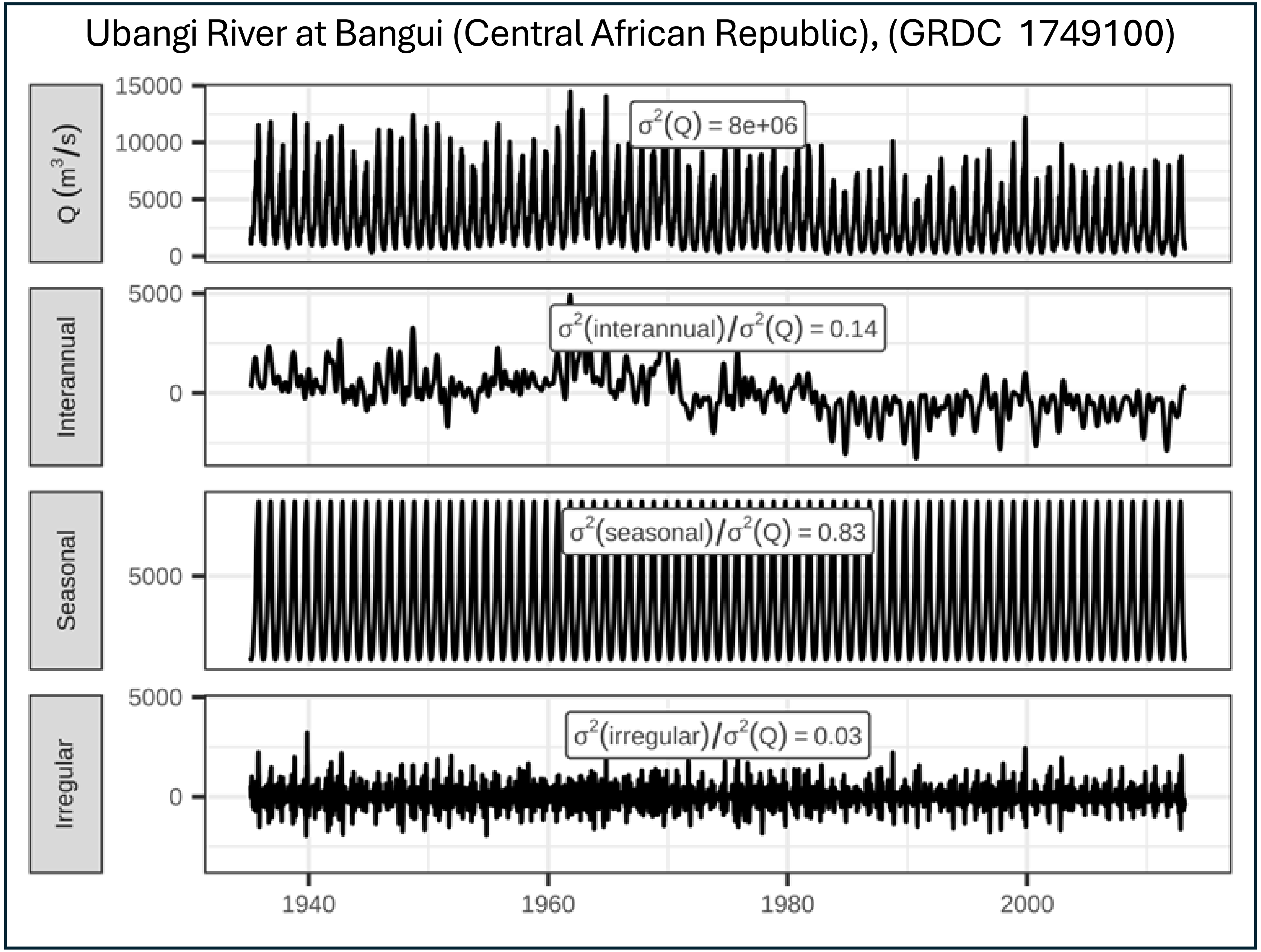


back to map:

Orinoco River at Puente Angostura (Venezuela), (GRDC 3206720)



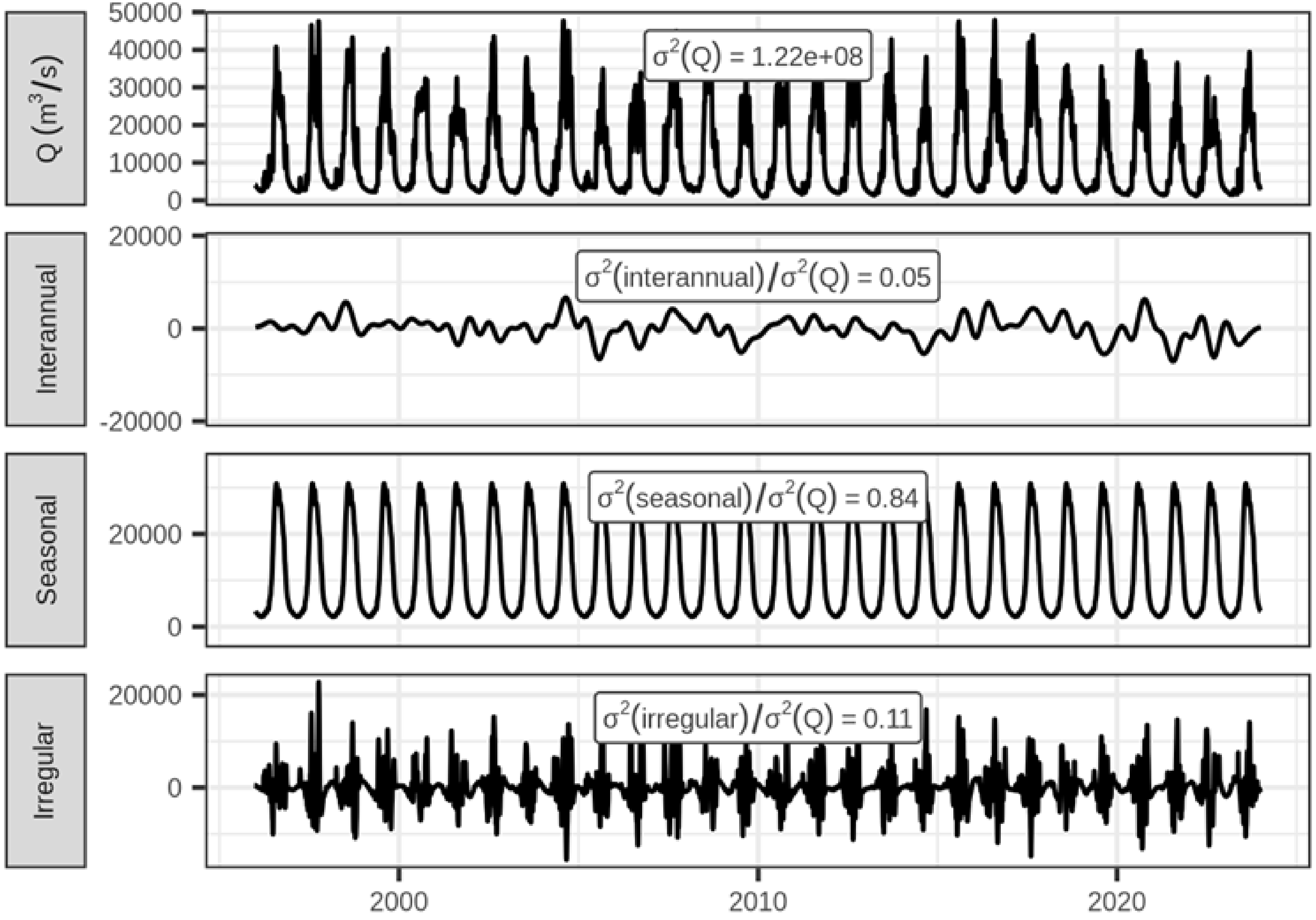
back to map:



Ubangi River at Bangui (Central African Republic), (GRDC 1749100)
Q (m³/s)
15000
10000
5000
0
σ²(Q) = 8e+06
Interannual
5000
0
σ²(interannual)/σ²(Q) = 0.14
Seasonal
5000
σ²(seasonal)/σ²(Q) = 0.83
Irregular
5000
0
σ²(irregular)/σ²(Q) = 0.03
1940
1960
1980
2000

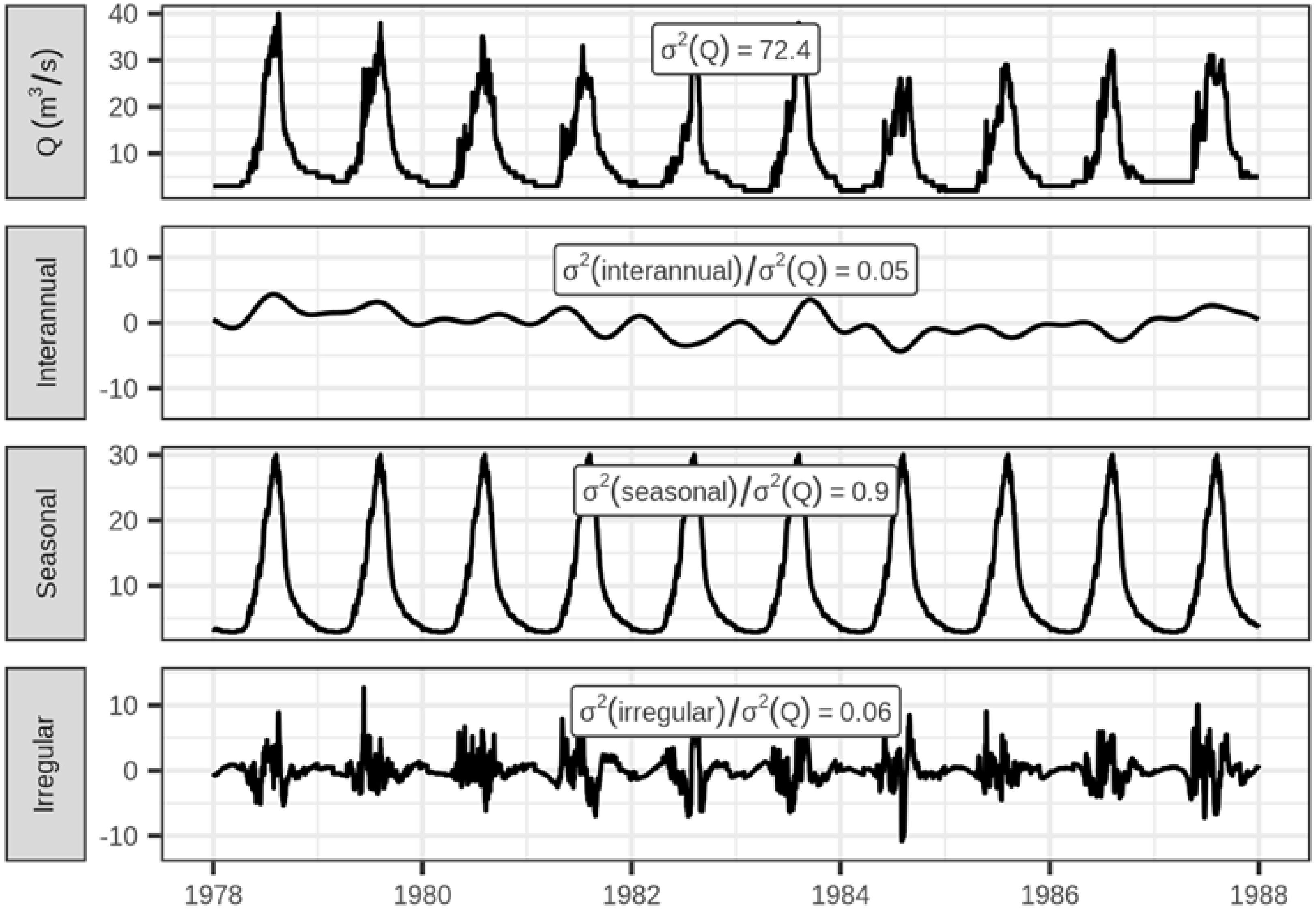
back to map:

Irrawaddy River at Pyay, Myanmar, (GRDC 2260700)

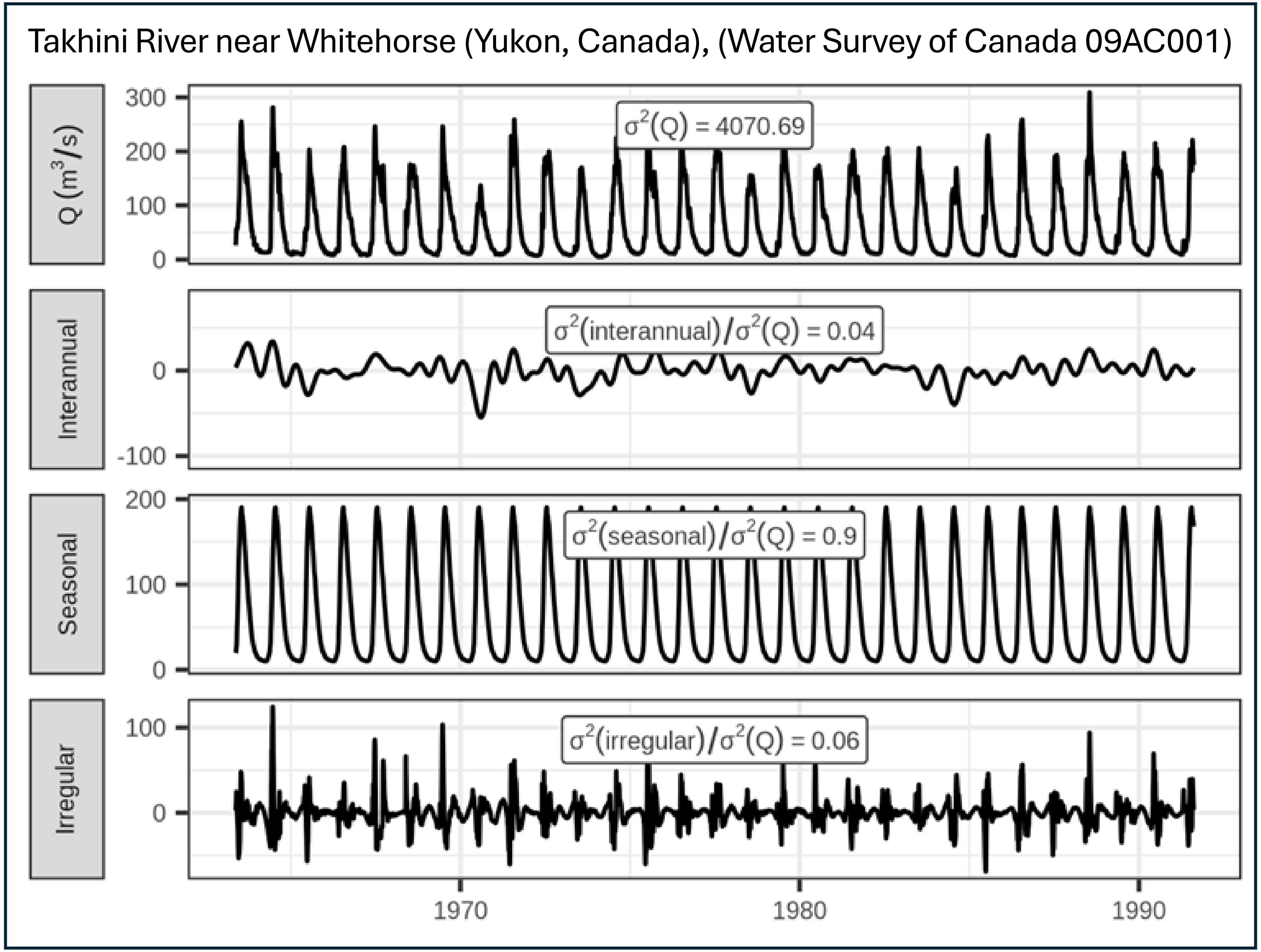


back to map:

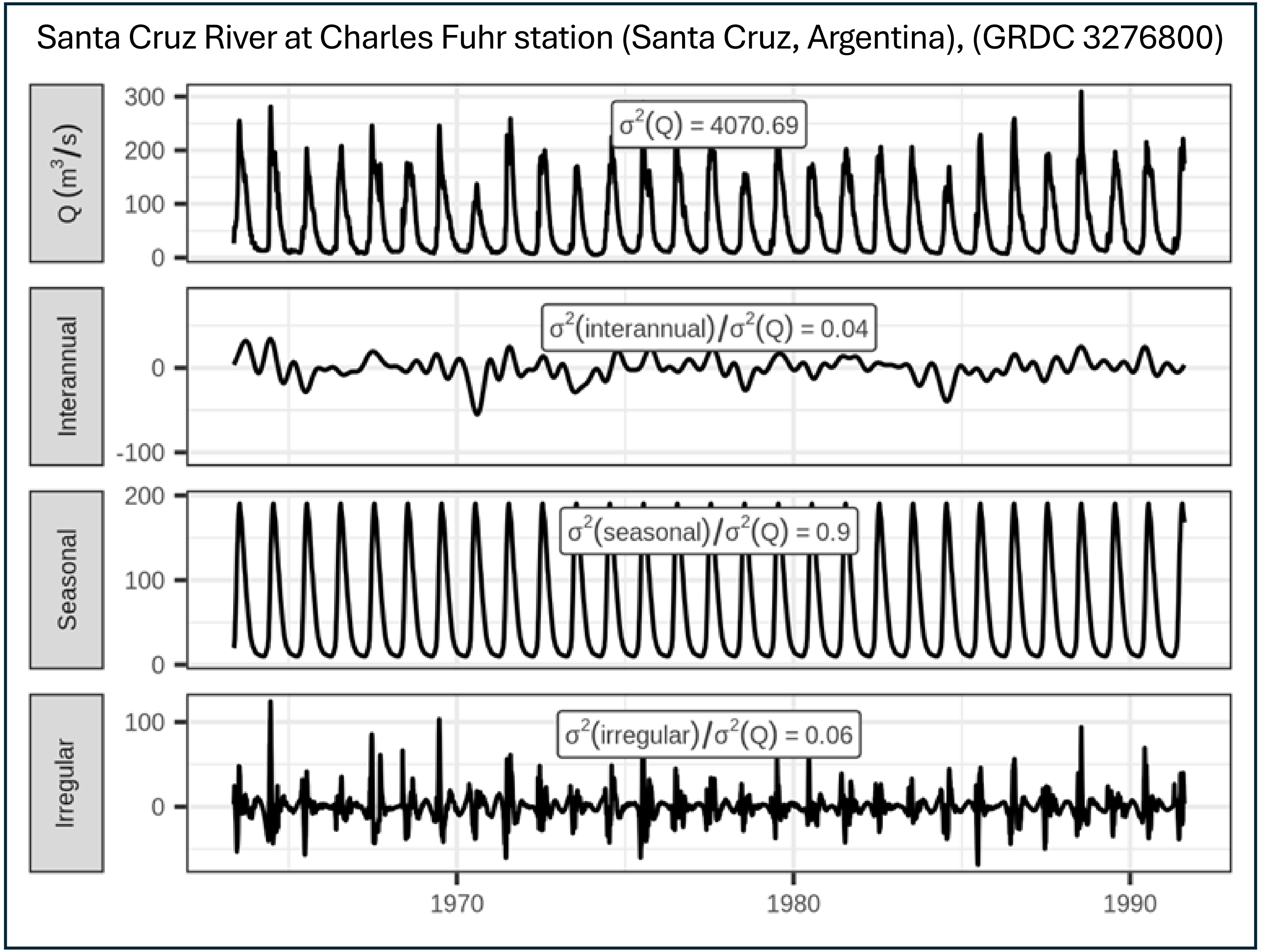
Talgar River at Talgar (Kazakhstan), (GRDC 2314400)



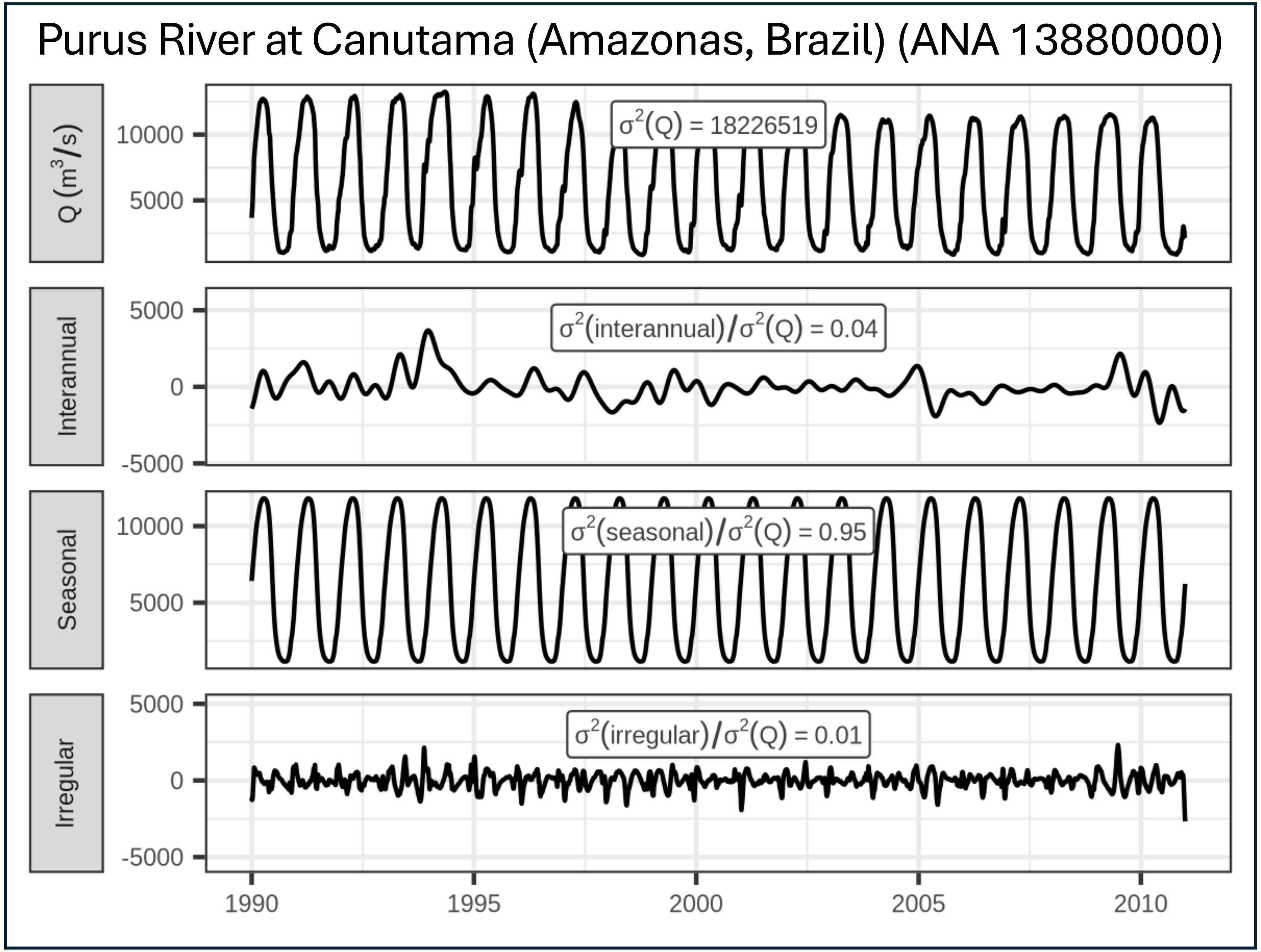
back to map:


Takhini River near Whitehorse (Yukon, Canada), (Water Survey of Canada 09AC001)
Q (m³/s)
Interannual
Seasonal
Irregular
$\sigma^2(Q) = 4070.69$
$\sigma^2(\text{interannual})/\sigma^2(Q) = 0.04$
$\sigma^2(\text{seasonal})/\sigma^2(Q) = 0.9$
$\sigma^2(\text{irregular})/\sigma^2(Q) = 0.06$
300
200
100
0
-100
1970
1980
1990

back to map:

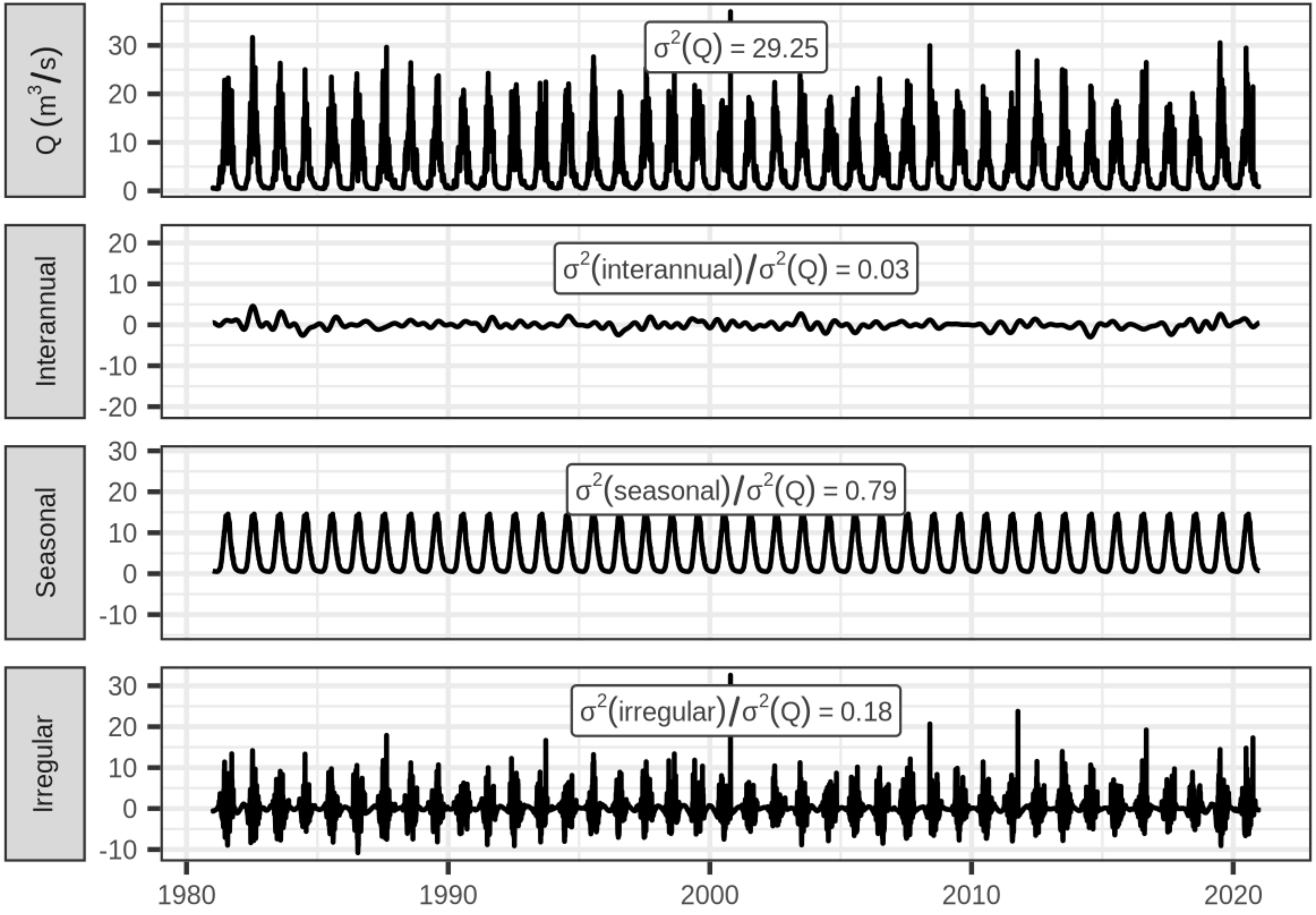

Santa Cruz River at Charles Fuhr station (Santa Cruz, Argentina), (GRDC 3276800)
Q (m³/s)
Interannual
Seasonal
Irregular
$\sigma^2(Q) = 4070.69$
$\sigma^2(\text{interannual})/\sigma^2(Q) = 0.04$
$\sigma^2(\text{seasonal})/\sigma^2(Q) = 0.9$
$\sigma^2(\text{irregular})/\sigma^2(Q) = 0.06$
300
200
100
0
-100
1970
1980
1990

back to map:


Purus River at Canutama (Amazonas, Brazil) (ANA 13880000)
Q (m³/s)
10000
5000
σ²(Q) = 18226519
Interannual
5000
0
-5000
σ²(interannual)/σ²(Q) = 0.04
Seasonal
10000
5000
σ²(seasonal)/σ²(Q) = 0.95
Irregular
5000
0
-5000
σ²(irregular)/σ²(Q) = 0.01
1990
1995
2000
2005
2010

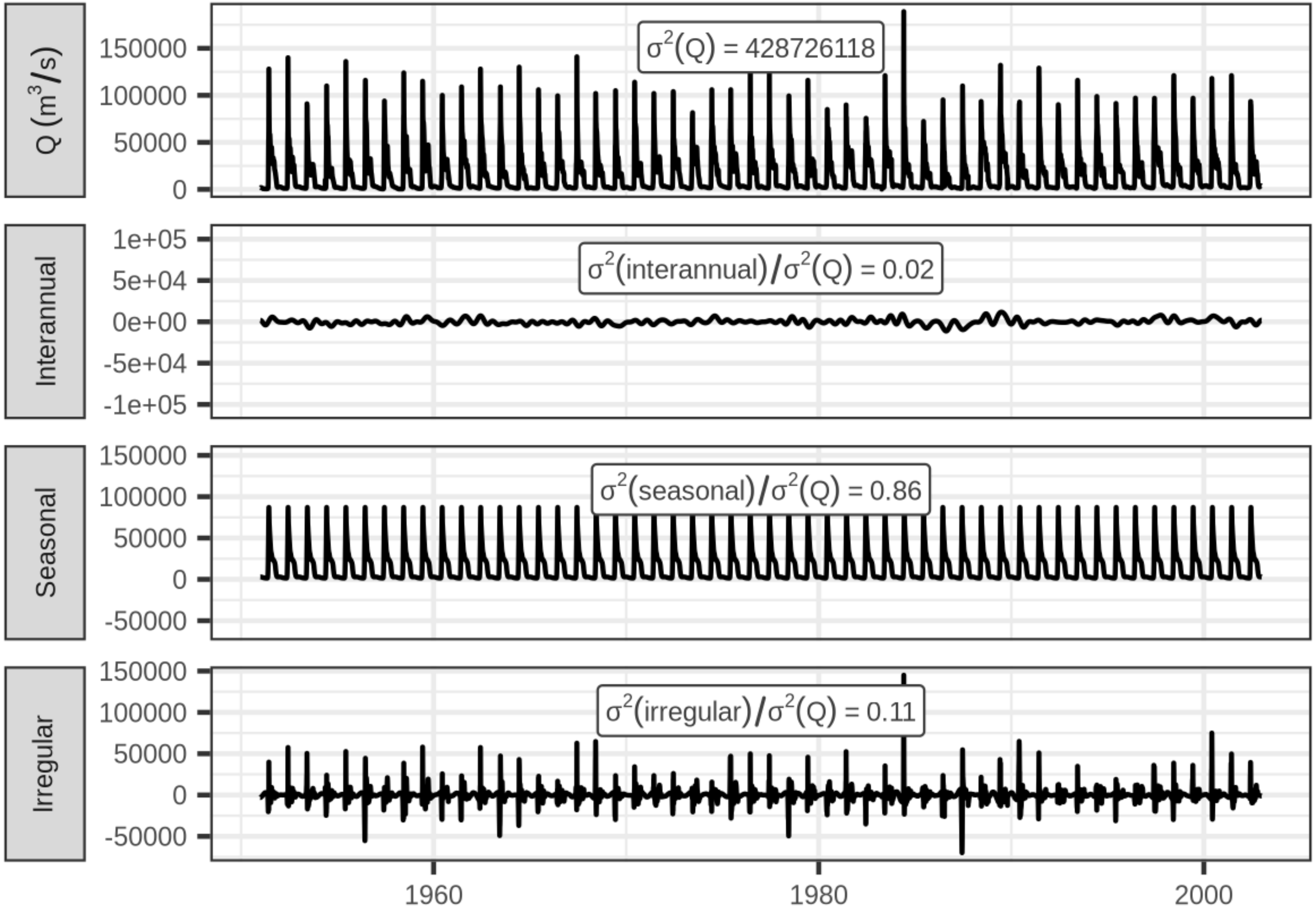
back to map:

Lonza River at Blatten (CAMELS-CH 2269)

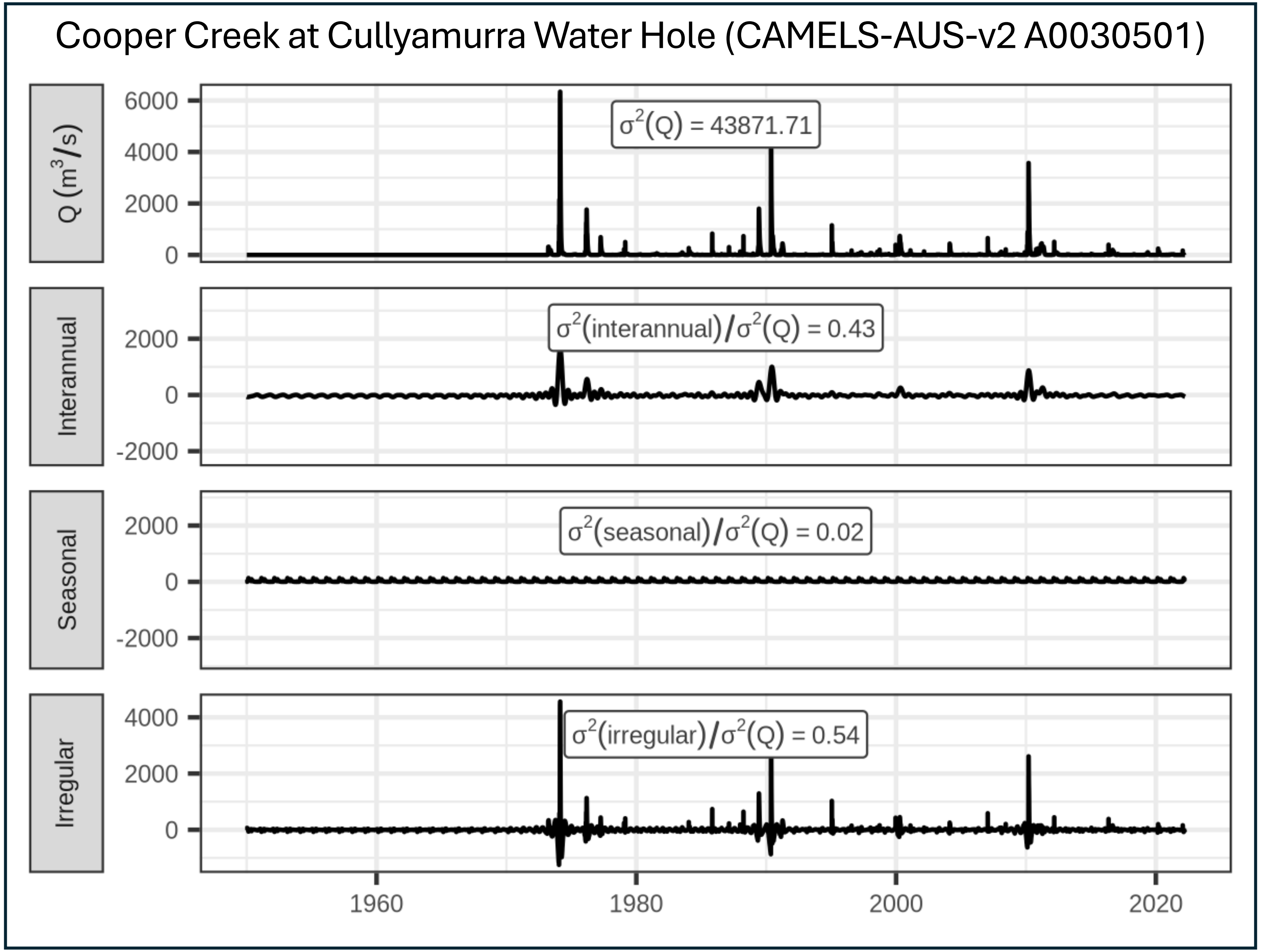


back to map:

Lena River at Stolb (GRDC 2903430)

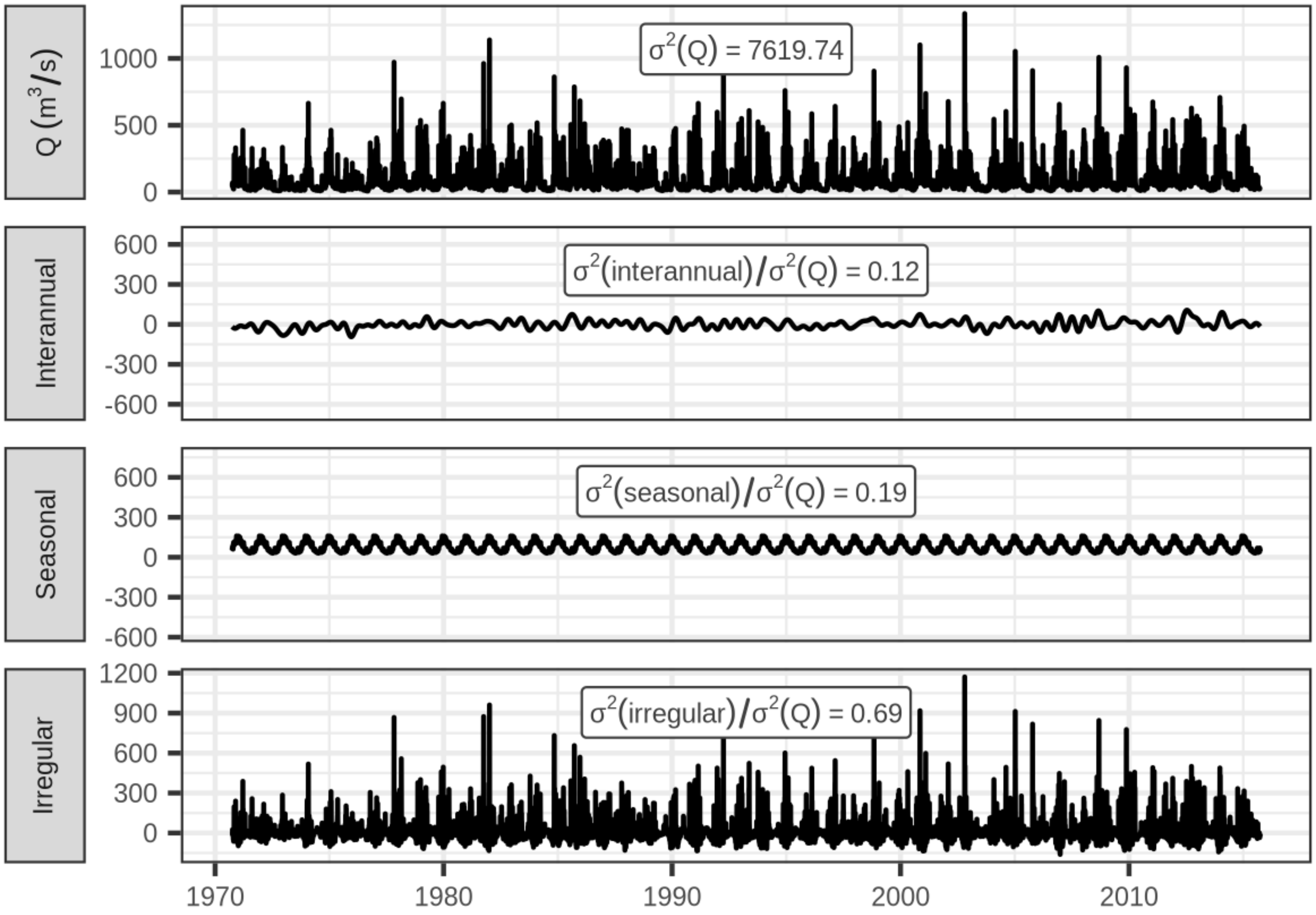


back to map:


Cooper Creek at Cullyamurra Water Hole (CAMELS-AUS-v2 A0030501)
Q (m³/s)
6000
4000
2000
0
σ²(Q) = 43871.71
Interannual
2000
0
-2000
σ²(interannual)/σ²(Q) = 0.43
Seasonal
2000
0
-2000
σ²(seasonal)/σ²(Q) = 0.02
Irregular
4000
2000
0
σ²(irregular)/σ²(Q) = 0.54
1960
1980
2000
2020

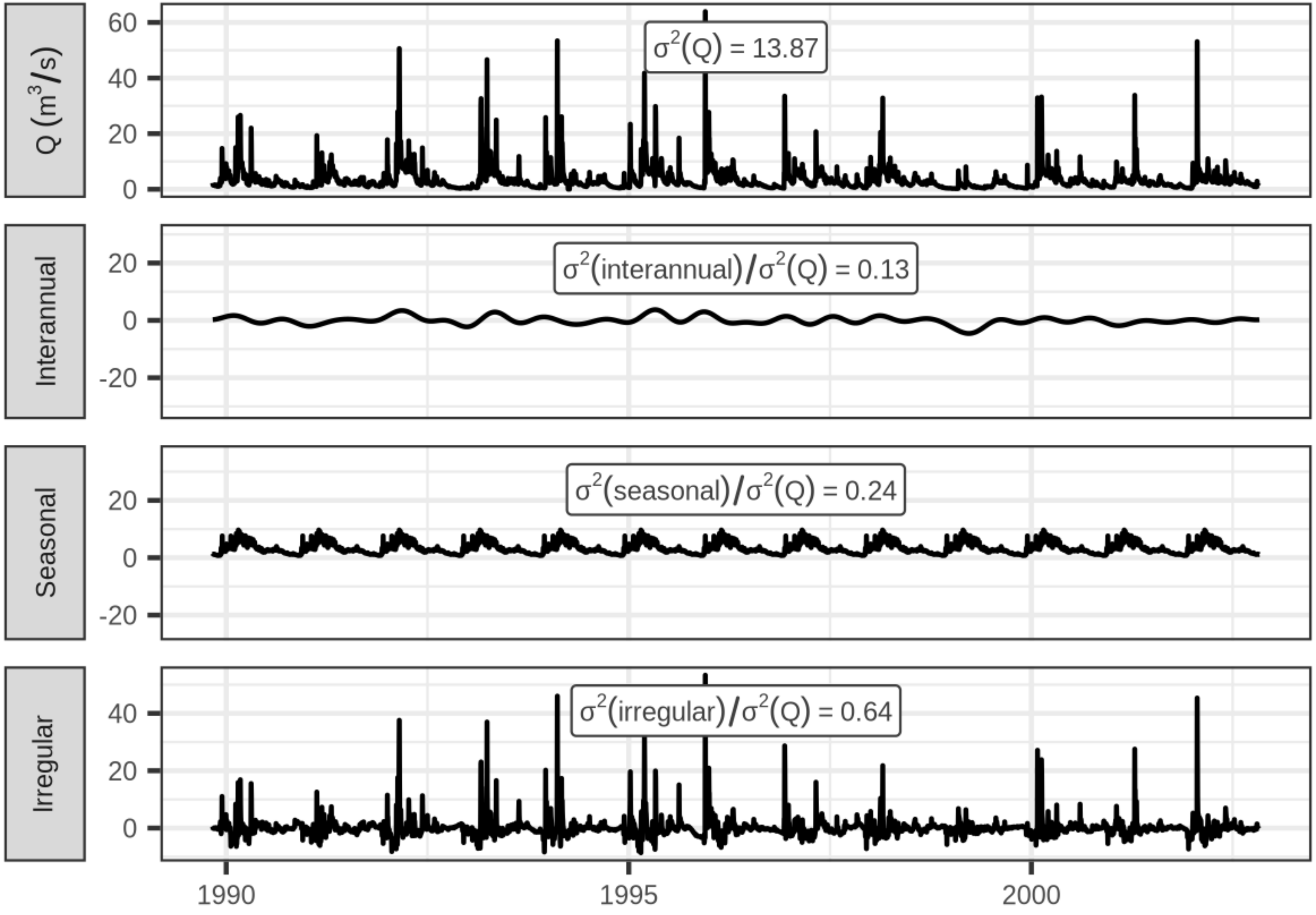
back to map:

Tweed River at Norham (CAMELS-GB 21009)

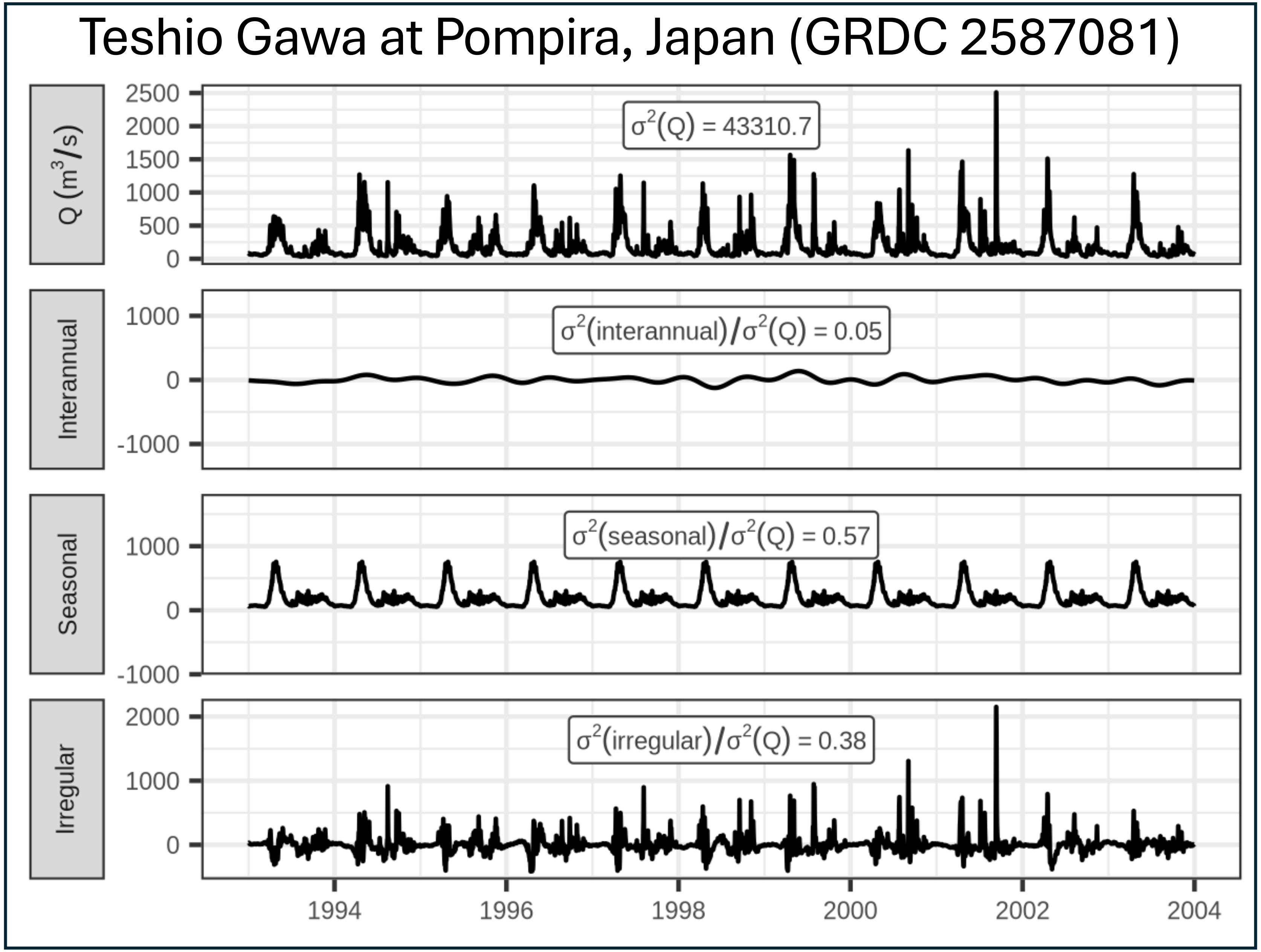


back to map:

RIVER DES CREOLES, Mauritius (GRDC 1689350)

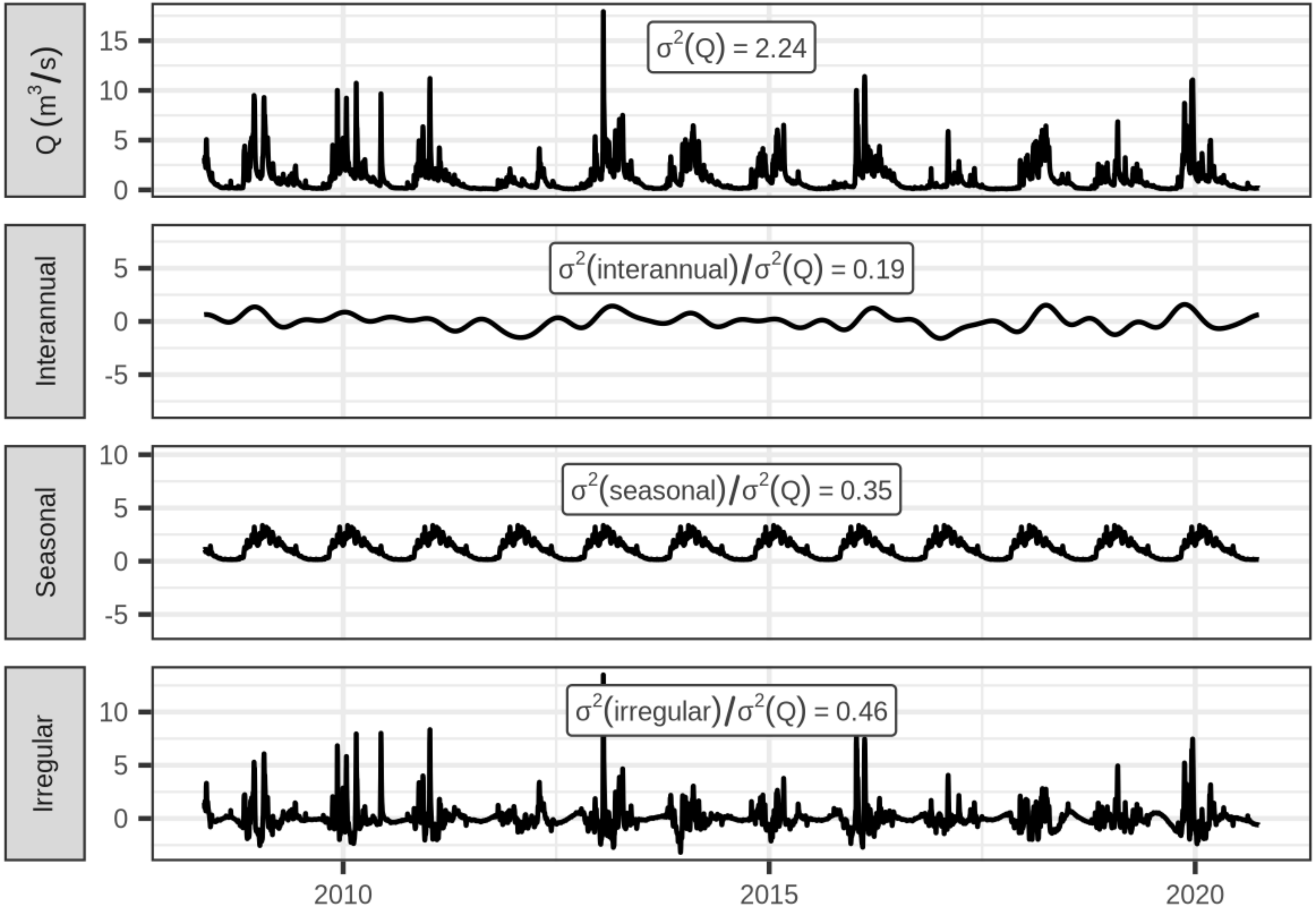


back to map:


Teshio Gawa at Pompira, Japan (GRDC 2587081)
Q (m3/s)
σ2(Q) = 43310.7
Interannual
σ2(interannual)/σ2(Q) = 0.05
Seasonal
σ2(seasonal)/σ2(Q) = 0.57
Irregular
σ2(irregular)/σ2(Q) = 0.38
1994
1996
1998
2000
2002
2004

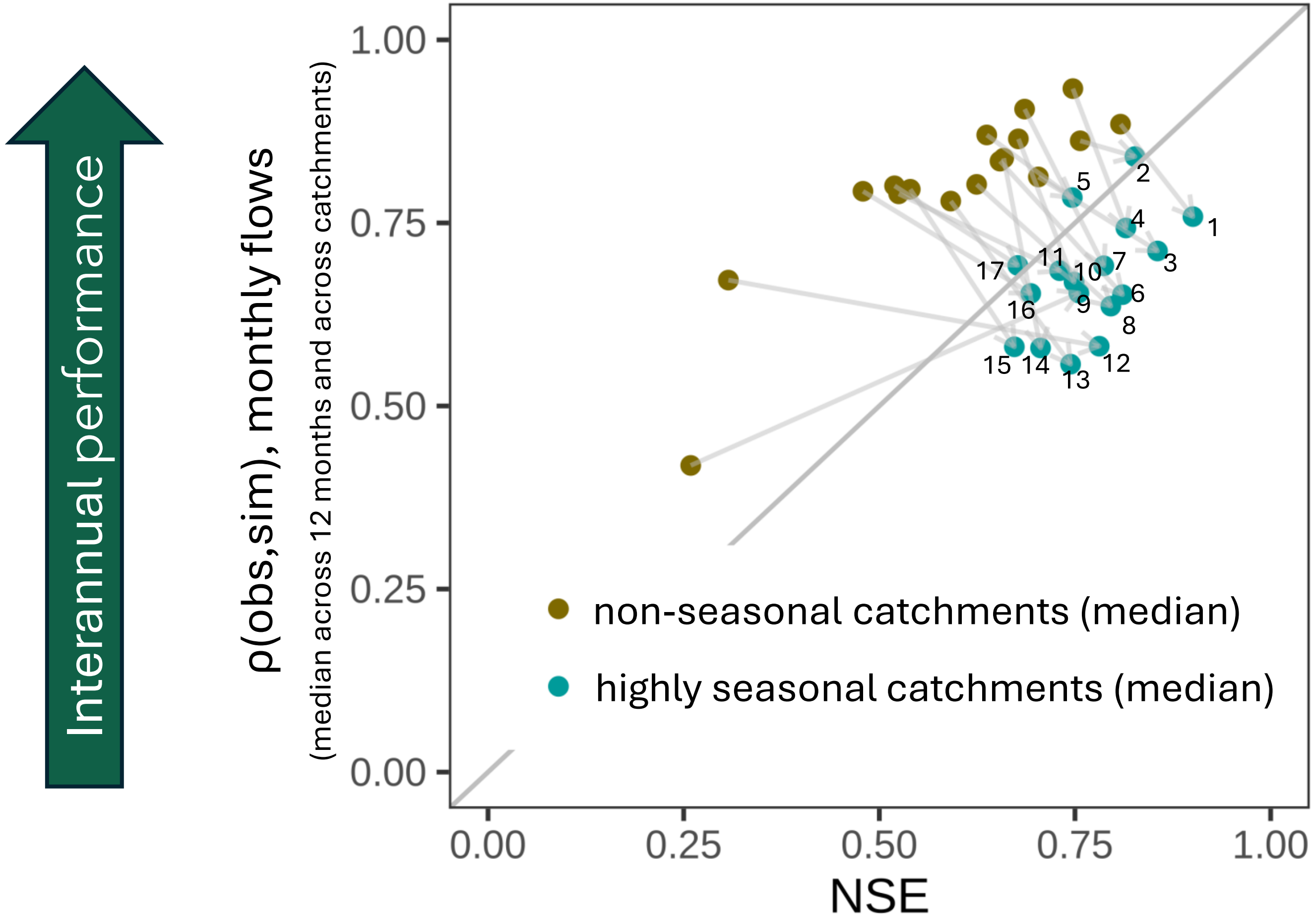
back to map:

Pontevilar, Galicia, Spain (CAMELS-ES 1605)



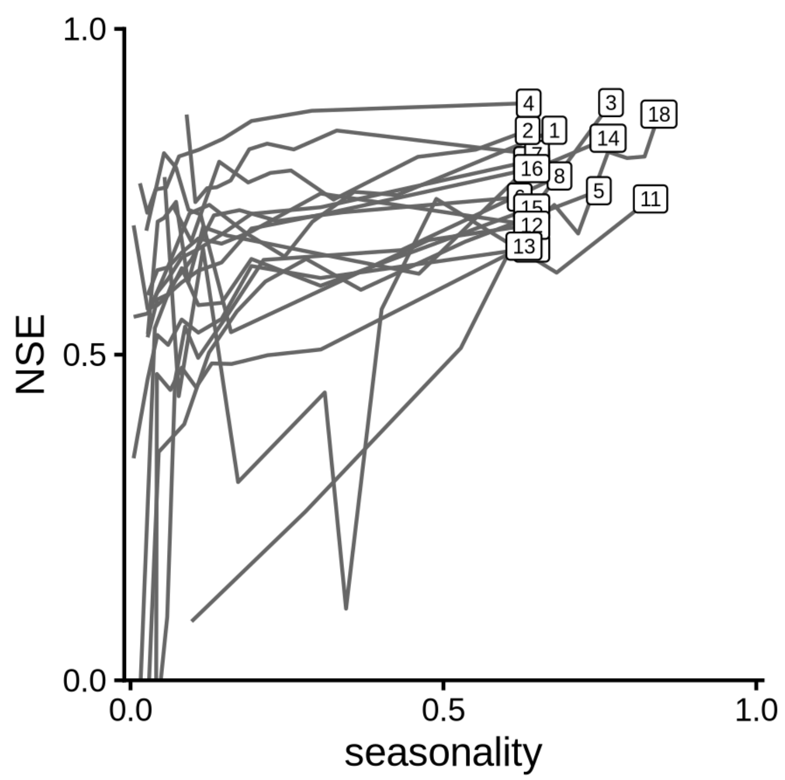
back to map:

Models achieve higher NSE in seasonal catchments but are worse at simulating interannual variability

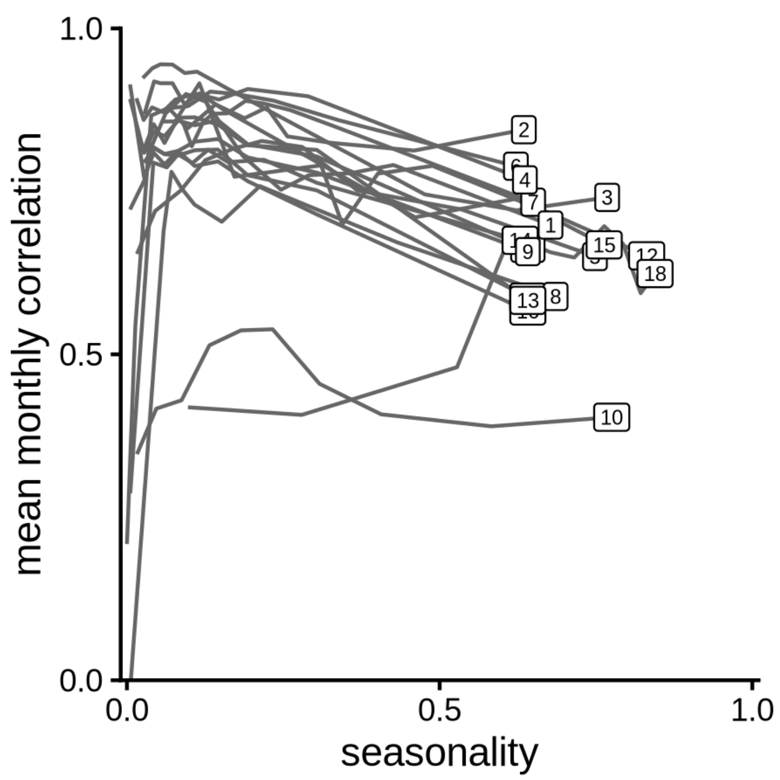


1) LSTM: Kratzert et al (2024), n = 531
2) LSTM: Arsenault et al (2022), n = 79
3) LSTM: Ruzzante et al (2026), n = 176
4) LSTM: Kraft et al (2025), n = 98
5) LSTM: Yang et al (2025), n = 3167
6) HBV: Seibert et al (2018), n = 671
7) PREVAH: Kraft et al (2025), n = 98
8) FUSE: Addor, Kratzert (2019), n = 576
9) VIC-GL (Schnorbus, 2018), n = 84
10) δHBV: (Song et al, 2025), n = 1131
11) LSTM: Nearing et al (2024), n = 3752
12) COSERO: Klingler et al (2021), n = 454
13) SAC-SMA: Newman et al (2017), n = 671
14) mHM: Mizukami et al (2017), n = 492
15) VIC: Newman et al (2017), n = 670
16) NHM: Regan et al (2019), n = 1340
17) MGB: Siquiera et al (2018), n = 33

Models achieve higher NSE in more seasonal catchments



… but are worse at predicting interannual variability



1) LSTM: Ruzzante et al (2026)
2) LSTM: Arsenault et al (2022)
3) LSTM: Kraft et al (2025)
4) LSTM: Kratzert et al (2024)
5) LSTM: Nearing et al (2024)
6) LSTM: Yang et al (2025)
7) δHBV: (Song et al, 2025)
8) COSERO: Klingler et al (2021)
9) FUSE: Addor, Kratzert (2019)
10) GloFAS (Nearing et al, 2024)
11) HBV: Seibert et al (2018)
12) MGB: Siquiera et al (2018)
13) mHm: Mizukami et al (2017)
14) NHM: Regan et al (2019)
15) PREVAH: Kraft et al (2025)
16) SAC-SMA: Newman et al (2017)
17) VIC: Newman et al (2017)
18) VIC-Gl: (Schnorbus, 2018)

Caravan-CMIP6

CMIP
Coupled Model Intercomparison Project

10 large-sample datasets

- Caravan
- CAMELS
- CAMELS-AUS-v2
- CAMELS-BR
- CAMELS-CH
- CAMELS-CL
- CAMELS-COL
- CAMELS-DE
- CAMELS-GB-v2
- CAMELS-IND

×

12 climate models

- ACCESS-ESM1-5
- CNRM-CM6-1-HR
- CanESM5
- EC-Earth3
- GFDL-CM4
- INM-CM5-0
- IPSL-CM6A-LR
- MIROC6
- MPI-ESM1-2-HR
- MRI-ESM2-0
- TaiESM1
- UKESM1-0-LL

×

Four scenarios

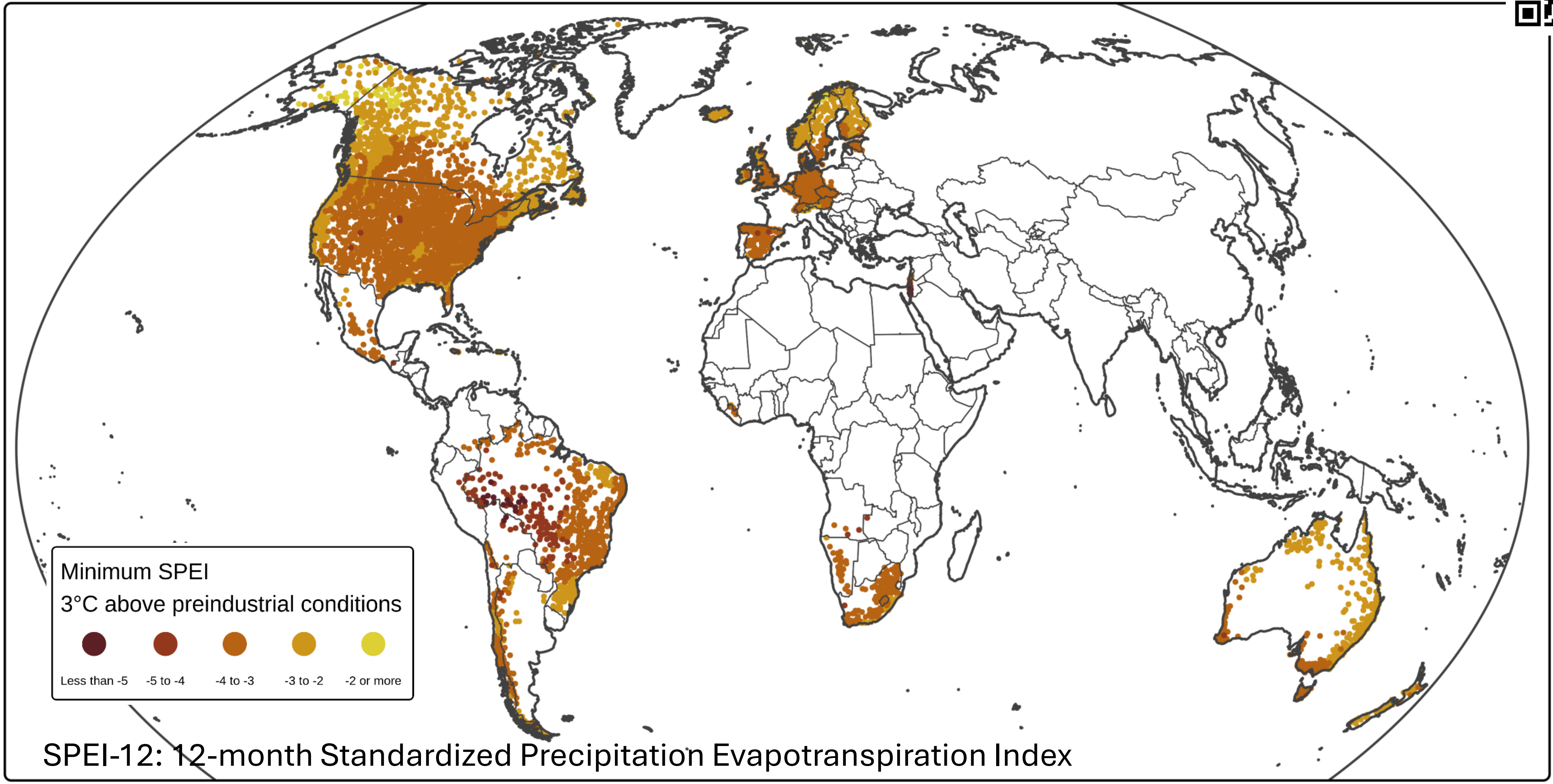
- Historical (1850-2014)
- SSP 1-2.6 (‘Sustainability’)
- SSP 2-4.5 (‘Middle of the road’)
- SSP 5-8.5 (Fossil-fueled Development)

×

Many variables

- Precipitation
- Temperature (min, max, mean)
- ET (various)
- Humidity (RH, vp, dewpoint)
- Radiation (net, shortwave, longwave)
- Air pressure
- Wind speed

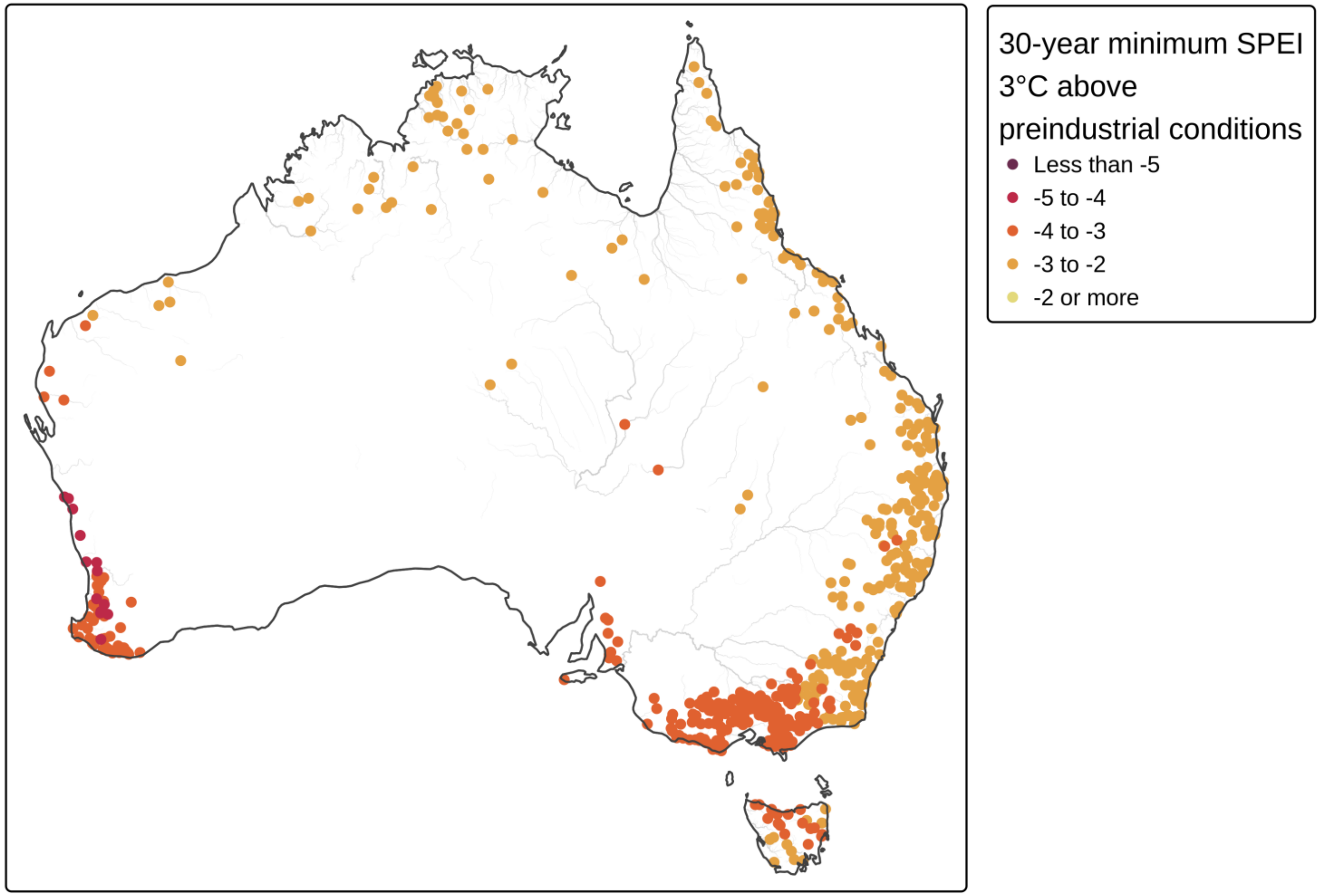
Caravan: SPEI-12 under 3^{0}C warming



SPEI-12: 12-month Standardized Precipitation Evapotranspiration Index

CAMELS-AUS-v2: SPEI-12 under 3^{o}C warming

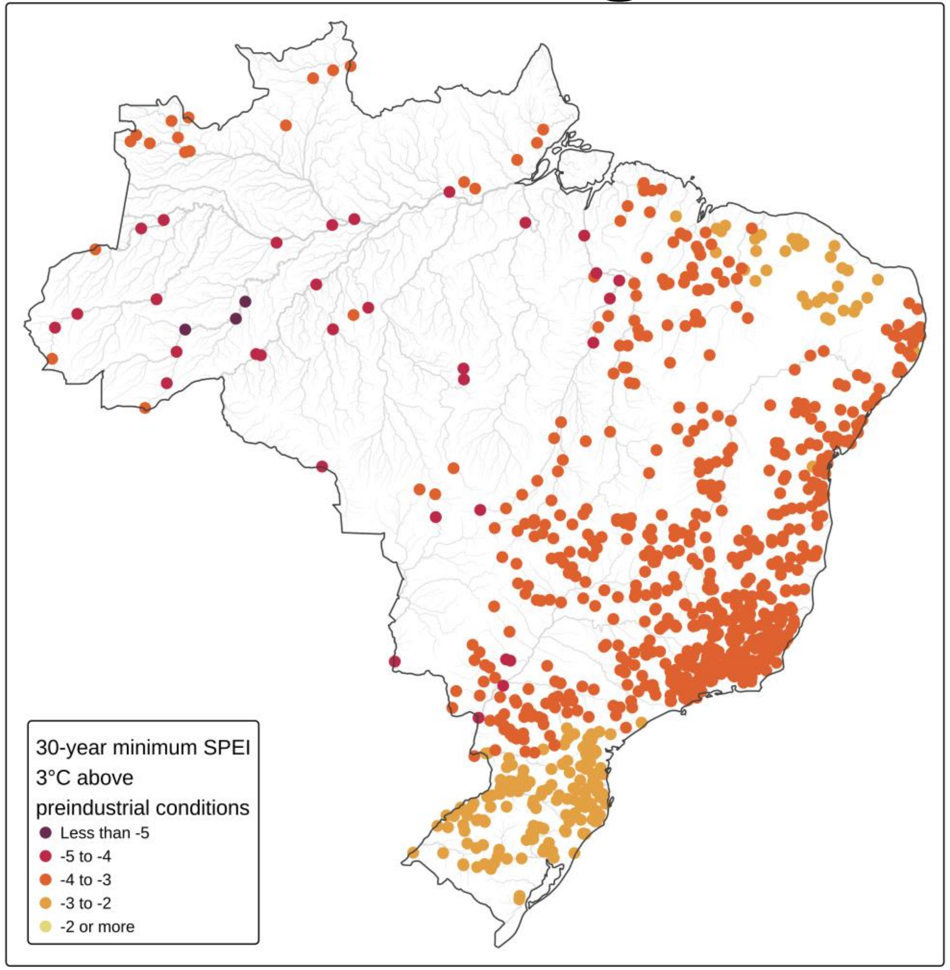
- Mostly SPEI [-3,-2) (return periods between 40 and 700 years)
- Victoria region mostly [-4,-3) (return periods between 700 and 30,000 years)
- Some catchments in Western Australia to see SPEI as extreme as -4.2 (return period of -70,000 years)



SPEI-12: 12-month Standardized Precipitation Evapotranspiration Index

CAMELS-BR: SPEI-12 under 3^{0}C warming

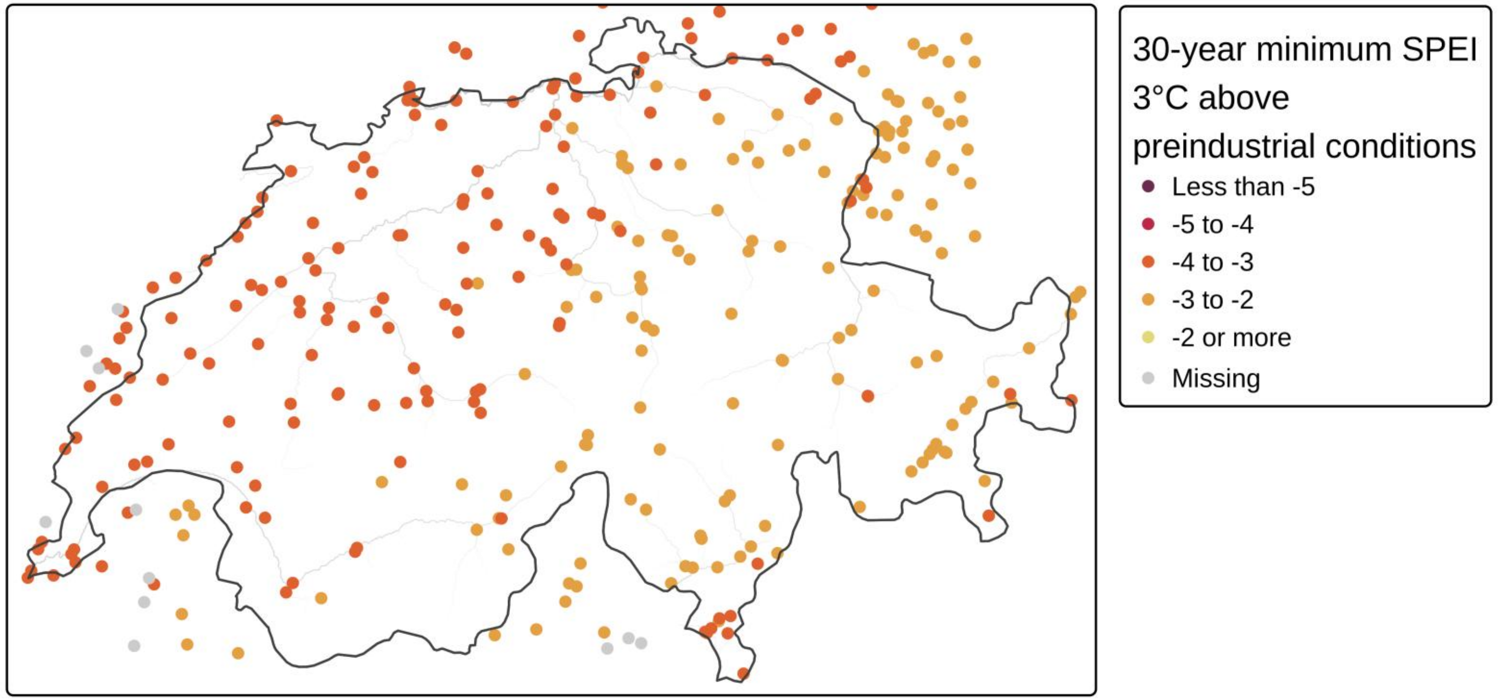
- Mostly [-4,-3) (return periods between 700 and 30,000 years)
- Amazon basin to values mostly [-5,-4), with most extreme values of -5.3 (return period of 2m years)



SPEI-12: 12-month Standardized Precipitation Evapotranspiration Index

CAMELS-CH: SPEI-12 under 3^0C warming

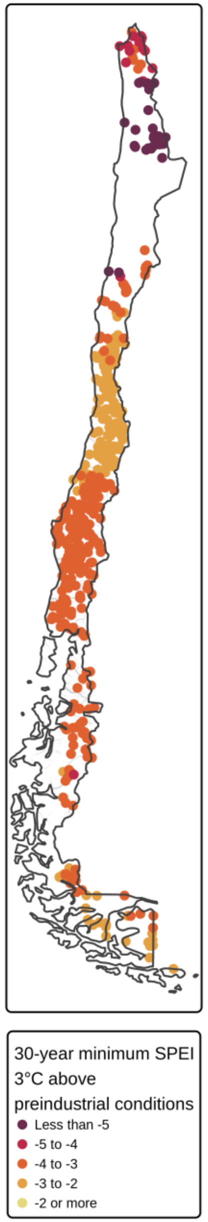
- Mostly [-4,-3) (return periods between 700 and 30,000 years)
- Less extreme droughts projected over alpine regions [-3,-2)



SPEI-12: 12-month Standardized Precipitation Evapotranspiration Index

CAMELS-CL: SPEI-12 under 3⁰C warming

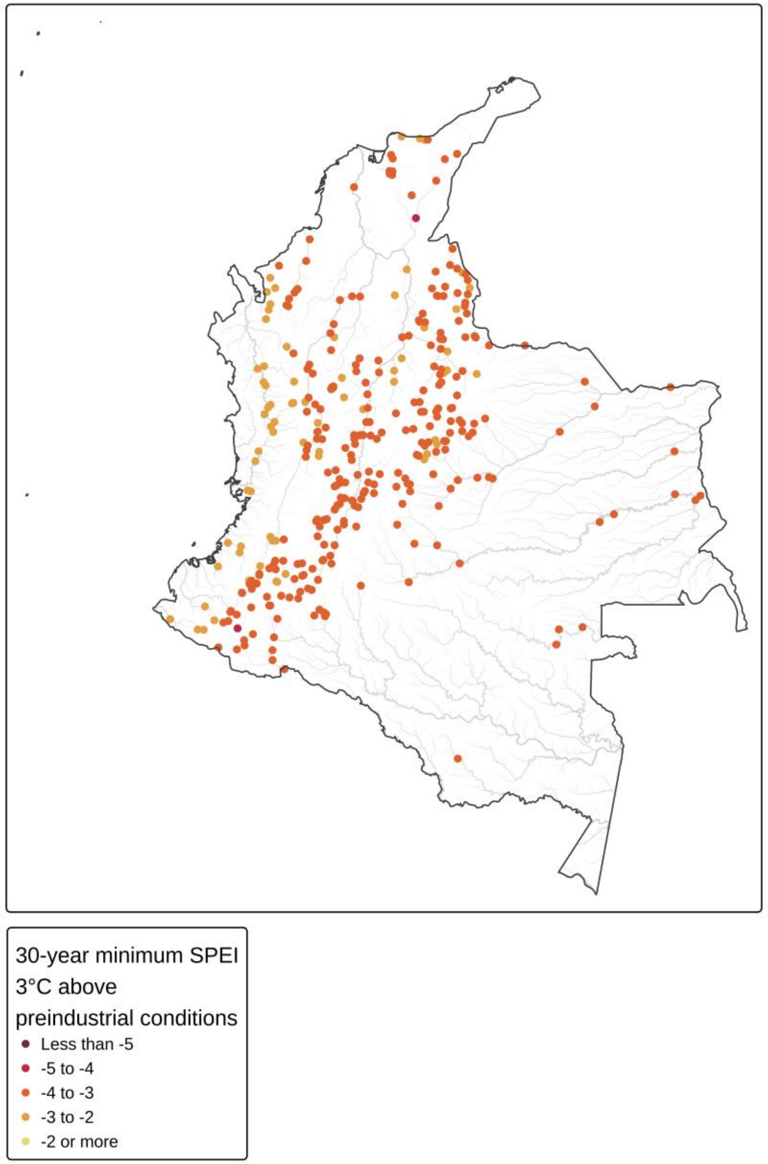
- Mostly [-4,-3) (return periods between 700 and 30,000 years)
- Less extreme droughts projected near Santiago [-3,-2)
- Most extreme droughts predicted in Atacama desert



SPEI-12: 12-month Standardized Precipitation Evapotranspiration Index

CAMELS-COL: SPEI-12 under 3^{0}C warming

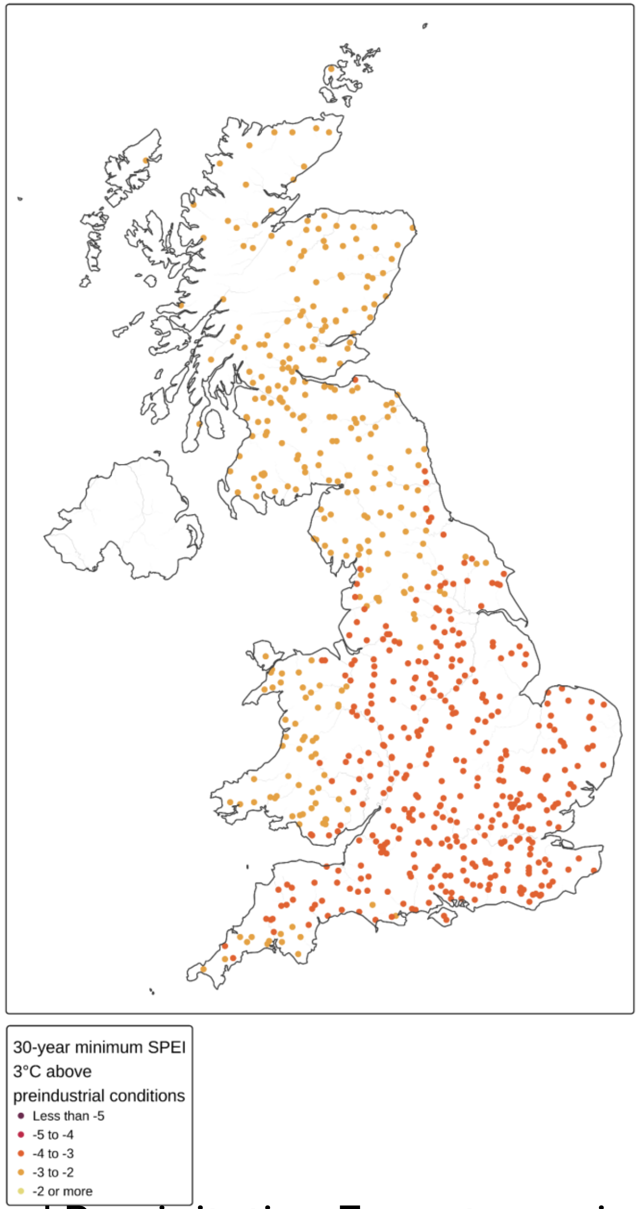
- Mostly [-4,-3) (return periods between 700 and 30,000 years)
- Less extreme droughts projected near Pacific coast [-3,-2)



SPEI-12: 12-month Standardized Precipitation Evapotranspiration Index

CAMELS-GB-v2: SPEI-12 under 3^{0}C warming

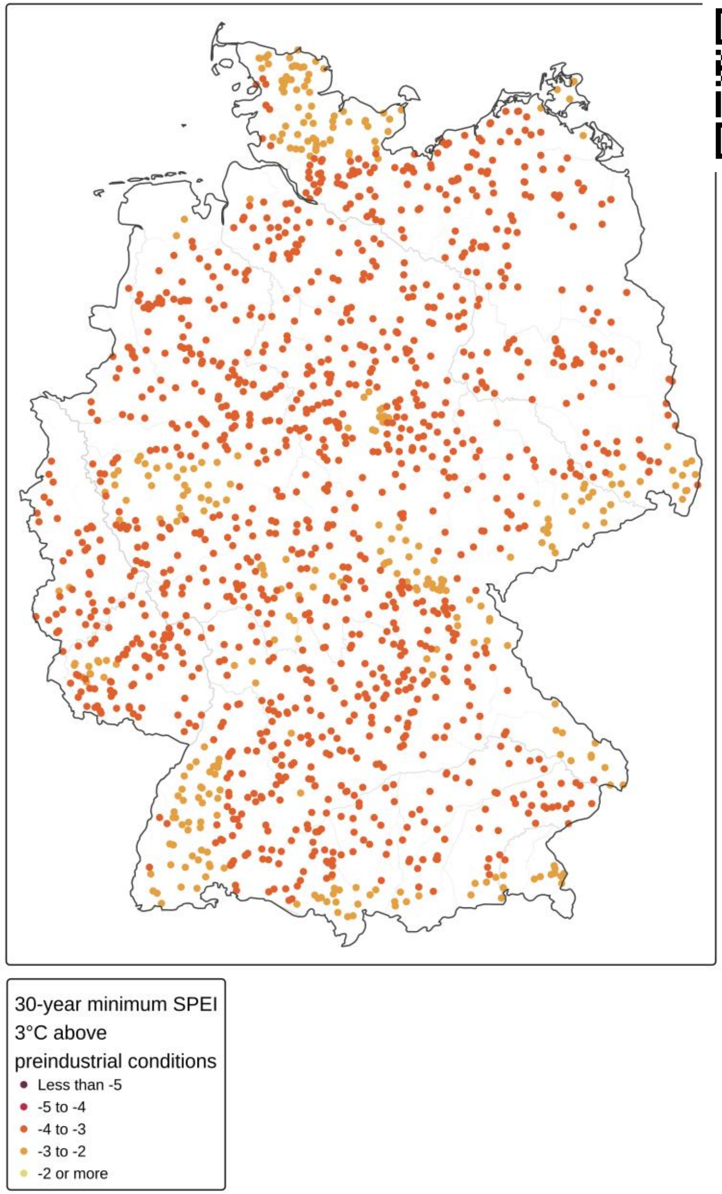
- Mostly [-4,-3) (return periods between 700 and 30,000 years) in southern England
- Somewhat less extreme in Wales, Northern England, and Scotland



SPEI-12: 12-month Standardized Precipitation Evapotranspiration Index

CAMELS-DE: SPEI-12 under 3^{0}C warming

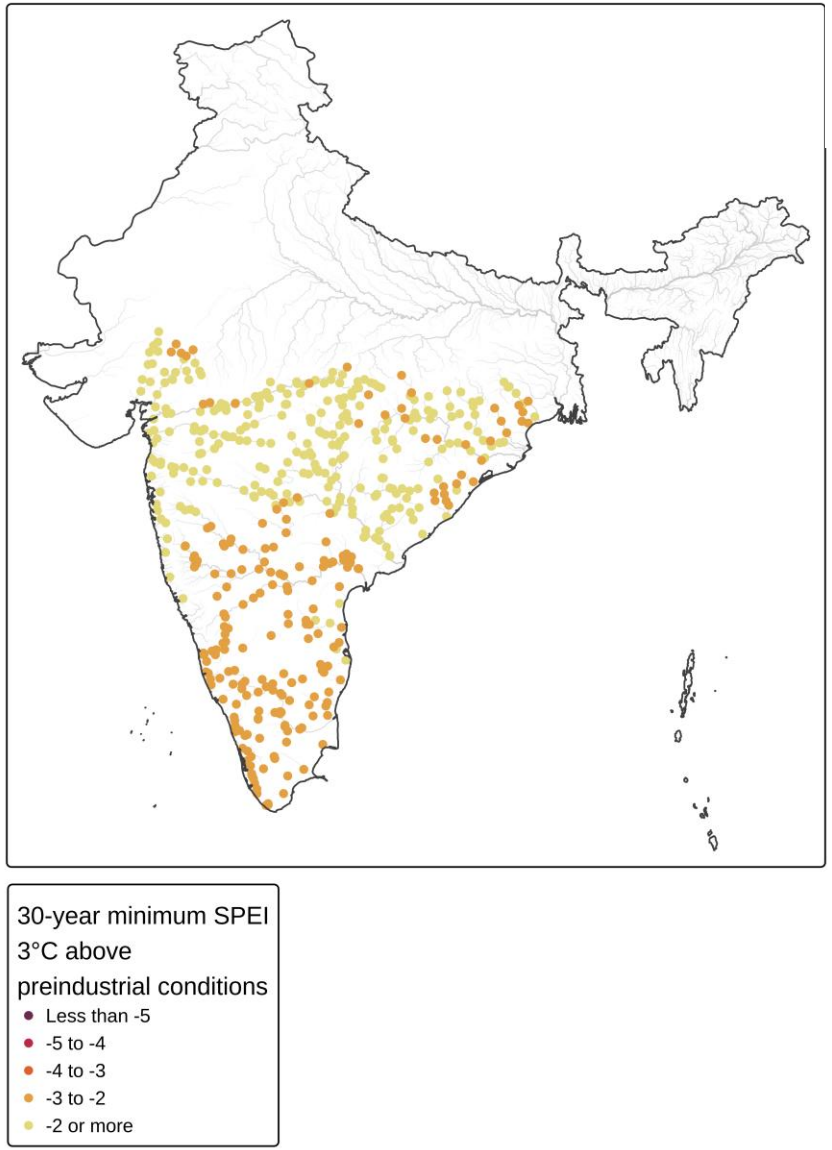
- Mostly [-4,-3) (return periods between 700 and 30,000 years)



SPEI-12: 12-month Standardized Precipitation Evapotranspiration Index

CAMELS-IND: SPEI-12 under 3⁰C warming

- Droughts expected to remain similar to historical conditions



SPEI-12: 12-month Standardized Precipitation Evapotranspiration Index