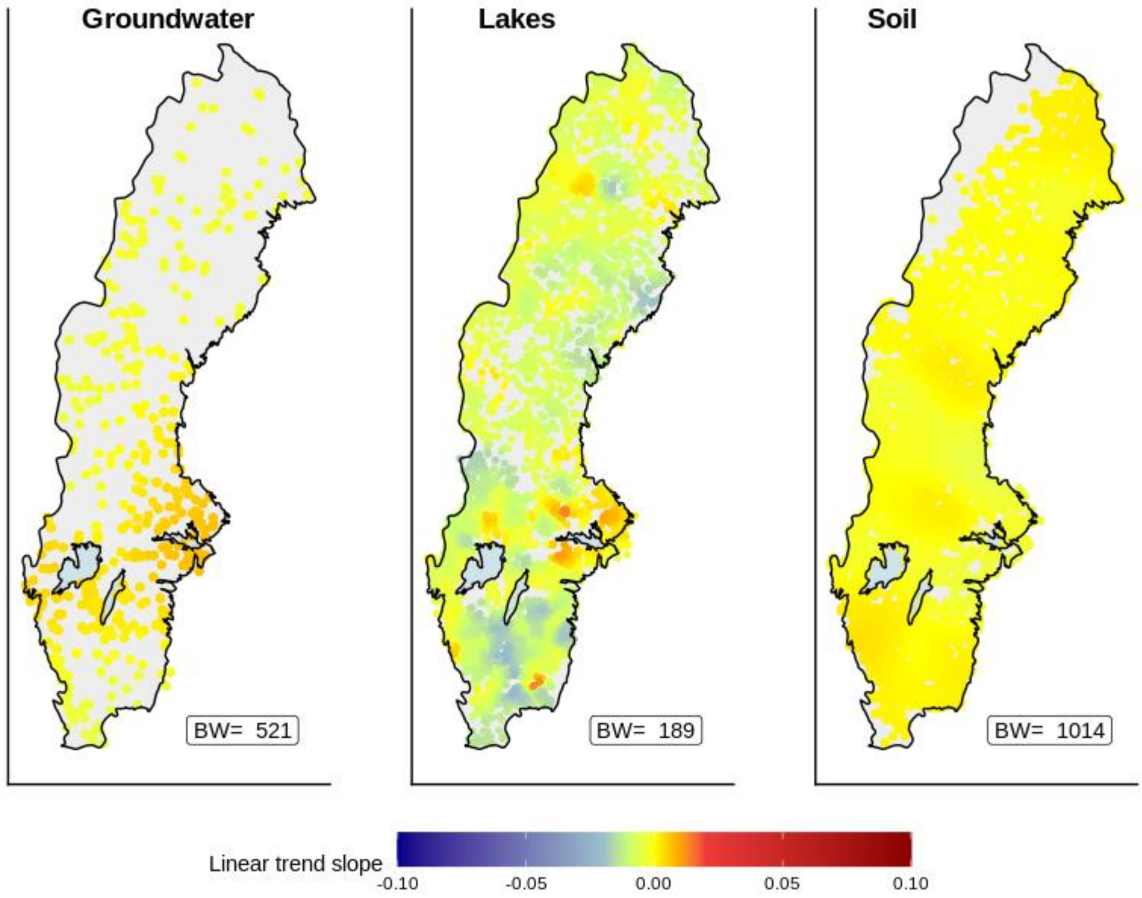
Identifying small- and large-scale drivers of trends in acidifying variables using multiscale geographically weighted regression models



Claudia von Brömssen, Maria Åkesson, Jens Fölster, Carina Josefsson Ortiz

Identifying small- and large-scale drivers of trends in acidifying variables using multiscale geographically weighted regression models

Claudia von Brömssen[1], Maria Åkesson[2], Jens Fölster[3], Carina Josefsson Ortiz[4]

1. Division of Applied Statistics and Mathematics, Department of Energy and Technology, Swedish University of Agricultural Sciences, PO Box 7032, 750 07, Uppsala, Sweden

2. Division of Groundwater Quality, Department of Geohydrology, Geological Survey of Sweden, PO Box 670, 751 28 Uppsala, Sweden

3. Section for Geochemistry and Hydrology, Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, PO Box 7050, 750 07, Uppsala, Sweden

4. Division of Biogeochemistry of Forest Soils, Department of Soil and Environment, Swedish University of Agricultural Sciences, PO Box 7014, 750 07 Uppsala

Report / Department of Energy and Technology, SLU
Report nr: 135
ISSN: 1654-9406

Department of Energy and Technology, SLU
Box 7032
750 07 Uppsala

https://www.slu.se/en/departments/energy-technology/
Uppsala, February 2026

Contents

Introduction

Monitoring programs designed for evaluation of trends rely on relatively frequent measurements in time but usually do not have spatial representation that is dense enough to connect them to explanatory variables that vary geographically. Other programs, like the so-called Omdrev-programs, monitor many objects, but with sparse temporal resolution and have therefore not been considered much in trend evaluations. In this report we evaluate the potential of geographically weighted regression models for trend analysis, including explanatory variables on three monitoring programs in Sweden: the Swedish Soil Inventory, the Swedish lake survey and the Swedish national environmental groundwater monitoring program, with measurements once each six or ten years for each of the objects.

Geographically weighted regression models (GWR, Brunsdon et al., 1998) have recently been adopted to allow the evaluation of regional trends (von Brömssen et al., 2023). They have also been tested to identify groups of observations which have similar levels or similar trends for a number of potential explanatory variables and exhibit similar trends in a response variable (von Brömssen et al., 2025). Both approaches rely on the principles of smoothing over windows of data which allows the use of not only the site of interest but also include sites that are either geographically or thematically similar. Since this approach allows an analysis even if the number of observations at a specific site is low, they are relevant for the analysis of trends in monitoring programs with low temporal but high spatial resolution. This report will focus especially on multiscale GWR (Fotheringham et al., 2017), since it is reasonable to assume that different explanatory variables work on different scale, e.g. that some might affect the response variable locally and others more regionally.

Trends in acidification were quite easy to evaluate during the 1980s and 1990s, where strong declines in acidifying variables were observed. Now, both the magnitude of trends in pH and SO_4 is smaller and both the magnitude and direction can vary substantially over Sweden (von Brömssen et al., 2021) due to several underlying natural processes and anthropogenic activities (Grennfelt et al., 2020). Temporal trends in pH could be influenced by reduced levels of sulphur deposition (Garmo et al., 2014; Vuorenmaa et al., 2018) as well as by changes in climate and increased eutrophication (Minella et al., 2013). Levels of pH in Sweden are generally low due to the high levels of natural organic acids and can vary due to stream flow conditions (Erlandsson et al., 2011).

For the current report we will investigate if a single explanatory variable can be added to GWRs additional to time and study its influence on trends in pH. We will:

- investigate the usefulness of both basic and multiscale GWR to identify effects of a driver variable on trends in pH
- determine if data collected in typical “omdrev”-surveys are sufficient to obtain reliable results from multiscale GWR
- evaluating pH trends in different media - soil, surface and ground water
- provide open-source R code that can be used to study additional variables in these monitoring programs.

Material and Methods

GWR

Geographically weighted regression models (GWR, Brunsdon et al., 1998) use geographical windows to provide a spatial smoothing over regression models. Typically, this is used to estimate spatially varying regression coefficients, i.e. to model if the relationship between two variables changes over space. It is straight forward to think of one of the explanatory variables in such as model as time. Then the goal of the analysis is to evaluate if the relation between a response and time changes over space and thus to determine if there are geographically varying trends for this variable. Some additional modification were made to this model to make trend evaluations as effective as possible (von Brömssen et al., 2023), which included station-wise mean centering of the response. The model is then

$$y_l^C = \beta_{l0} + \beta_l \cdot t + \varepsilon_l$$

where y^C indicates station-wise mean centered values of the response, l denotes the geographical location and t_l is a time variable, typically the year of observation. β_l represents the regression coefficient, i.e. the change in the response per time unit, which is dependent on location l.

This model is estimated within several geographical windows. Within each geographical window the observations are weighted according to distance to the centre of the window to obtain a smooth transition over space. For this weighting often a bi-square weight is used. The bandwidth in GWR, which corresponds to the window-size, can be set as k-nearest-neighbours (knn, adaptive) or as distance based (non-adaptive). Using knn leads to windows that always contain the same number of observations, while distance-based windows include an area of a certain radius independently of how many observations are within this area. The optimal bandwidth for a certain analysis is determined by cross-validation, i.e. a search for a bandwidth that produces the best predictions is conducted. In this study we use the knn-approach and the optimised bandwidth is presented in combination with the different results.

The basic GWR model can be extended to include one or a few explanatory variables. With one explanatory variable the model is

$$y_l^C = \beta_{l0} + \beta_{l1} \cdot t + \beta_{l2} \cdot x_{l2}^C + \varepsilon_l$$

including the station-wise mean centered variable x_{l2}^C. In the basic model the same bandwidth is used for the time variable and the other explanatory variable. This means, that the size of the window is the same for both variables.

Multiscale GWR

Multiscale GWR (Fotheringham et al., 2017) is a version of GWR that allows the bandwidth to be different for included explanatory variables. This can lead to the identification of a global or regional effect for one variable, while the relationship to another variable is more locally defined. This allows the model to distinguish between large-and small-scale drivers, which can be important if underlying processes are to be understood.

Software

Computations were made in the open-source software R (R Core Team, 2025) using the package GWmodel (Gollini et al., 2015). In order to make multiscale GWR more attractive for implementation with large datasets, as in this case the lake data, we improved the code by reducing memory copies (Höglund and von Brömssen, 2025), which was added to the version of GWmodel publicly available. The new version is considerably faster. However, larger models do still use a lot of the computer's working memory, making the computer unavailable for other tasks. For this study, computations were made on SLU's R Studio server, which is a virtual Linux-server with similar capacities as the author's PC.

Data

The National environmental groundwater monitoring program

The national environmental groundwater monitoring program has been revised several times since its origin in the 1970s. For the past c. 20 years, since the implementation of the Water Framework Directive, the program has run on a six-year cycle that aims at covering in total about 600 sampling sites per cycle. Around 80 of the 600 sites are so called trend stations that are sampled either 2 (spring, late summer/early autumn) or 4 times per year. The trend stations are complemented by circulation sites (omdrevsstationer), that have a revisit time of six years, i.e. about 80-100 circulation sites are sampled in August every year. Sampling sites include primarily natural springs and observation wells, but also some private and municipal drinking water wells. The program aims at monitoring primarily relatively shallow groundwater environments with low turnover times, in order to be able to pick up effects of external pressures relatively fast. Sampling sites, both trend and circulation sites, are distributed all throughout the country over the full six-year cycle, although circulation sites tend to be somewhat clustered geographically each year due to logistic purposes. In this study, all circulation sites were included using observations between 2007 and 2024.

The Swedish lake survey

The national monitoring program the Swedish Lake Survey (SLS) includes approximately 4800 lakes larger than 0.01 km^2 from the national lake register within geographical and lake size strata (Fölster et al., 2014a, 2014b). The programs have a revisit time of six years, meaning that a single station has one observation every six years and each year about one sixth of the lakes are monitored. The lakes sampled within a specific year are similarly distributed over the country, but the monitoring density is generally higher in southern Sweden. Mid-lake surface samples are taken during autumn circulation, which usually occurs between September and December, since these measurements are considered to be representative of the entire lake (Göransson et al., 2004). Occasional late observations (up to January) are attributed to the autumn of the previous year. During the autumn of 2018 and January 2019 fewer lakes were sampled due to the technical breakdown of the helicopter that is used for their monitoring. The remaining lakes were sampled together with the panel planned for 2019. Certain data quality problems were detected for the first year of monitoring in the SLS (2007), therefore only data from 2008 to 2024 was included in this study.

The Swedish forest soil inventory

The Swedish Forest Soil Inventory (SFSI) is a repeated national soil inventory with around 20 000 plots covering the whole country. The SFSI has been integrated with the Swedish National Forest (NFI) Inventory since 1983. The inventory is made on all land except cropland, settlements and high mountains. The inventory has a 10-year cycle, and 1/10 of the plots are inventories distributed over the country each year. The design is formed in square clusters of the plots. The size and distribution of the clusters vary over the country, with larger and fewer clusters in the north. Only half of the plots are used within the SFSI. On each inventoried plot, a soil description is made and samples collected for analysis of chemical variables such as C, N, pH, cat ions, exchangeable aluminum and DNA.

The selection of plots in this study was from the third inventory between 2003 and 2023, and the variables used in this study were land use, soil moisture, humus form, humus depth, soil depth, parental material, texture, soil type, C, N, pH, Ca, K, Mg, Mn, Na and TA.

Selecting appropriate explanatory variables and evaluating model results

In the GWR and multiscale GWR context one or several explanatory variables can be included. However, especially for multiscale GWR, the computations get very computationally heavy. Therefore, in this work we concentrate on only one explanatory variable.

Which variable should be selected depends on the research question. Here, we base the selection of the explanatory variable on publication with similar purpose (von Brömssen et al., 2025), where a number of chemical variables, land use variables and climate related variables were included to explain lake pH trends. The variable most closely positively related to pH trends were changes in calcium (Ca). A similar analysis was made for groundwater (Supplementary, Figure S1), which also indicates a relationship between trends in pH and Ca. However, in this case the relationship was negative and not as strong as for lake surface water. In this case study we will use Ca as an explanatory variable explaining the level of pH, while simultaneously estimating the temporal trend in pH. Both pH and Ca are station-wise mean-centred and Ca is also log-transformed prior to any computations. Observations on Gotland were removed for this study, due to the specific characteristics compare to and long distance from other observations.

When an explanatory variable is added to a GWR model we can evaluate the results in different ways:

- summarize and visualise the (spatially varying) coefficients, i.e. the relationship between the response and the selected explanatory variable
- compare temporal trends in the response variable using the model with and without the explanatory variable included

These evaluations will be done for both basic and multiscale GWRs and for lakes, ground water and soil.

Results

Geographically weighted regression models for trend in pH using only time as explanatory variable

When no explanatory variable was included models for groundwater and lakes indicated some trends with small magnitude in the Stockholm area and – for lakes – in the South-West of Sweden (Figure 1). For both these models the bandwidth was determined to be about 800, but since the bandwidth indicates the k-nearest-neighbors the smoothing was more substantial for the groundwater dataset (with fewer observations in total). For the soil dataset the bandwidth was high, indicating a global smooth, i.e. no local trends could be identified. There was also no overall trend in pH for this dataset. For comparisons trends for Ca are shown in Figure S.2.

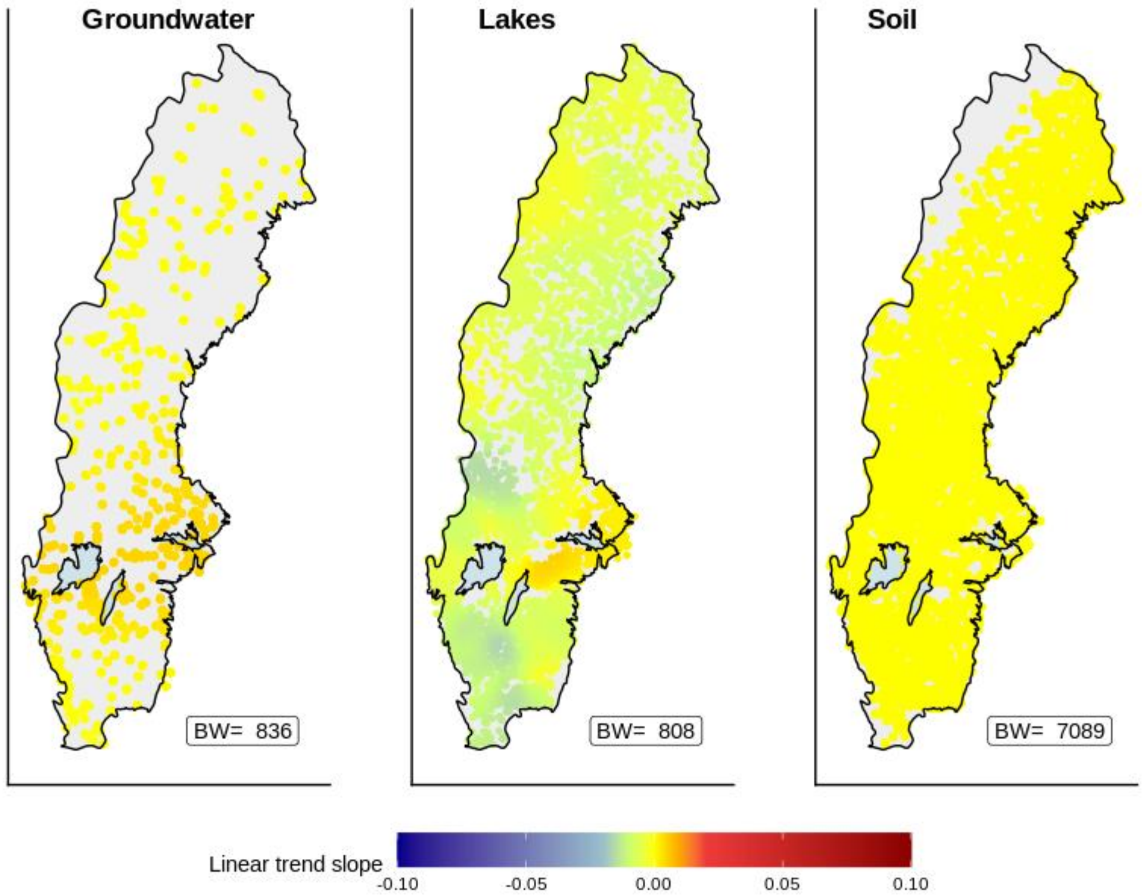


Figure 1: Geographically weighted trends for pH. Left: Groundwater (2007-2024), Middle: Lakes (2008-2024), Right: Soil (2003-2023)

Basic geographically weighted regression models for trend in pH using time and calcium as explanatory variable

Including Ca as an explanatory variable we studied both the spatially-varying effect of this variable on the response and which trends were visible after adjusting for changes in Ca. The trends were generally stronger in the adjusted data (Figure 2), where trends with higher magnitude were observed in the Stockholm area (positive trends) for groundwater and lakes. For lakes the negative trends in South-West Sweden persisted. For soil the trends observed were still weak and the smoothing was higher than for the other media. Some positive trends were observed in South-West Sweden.

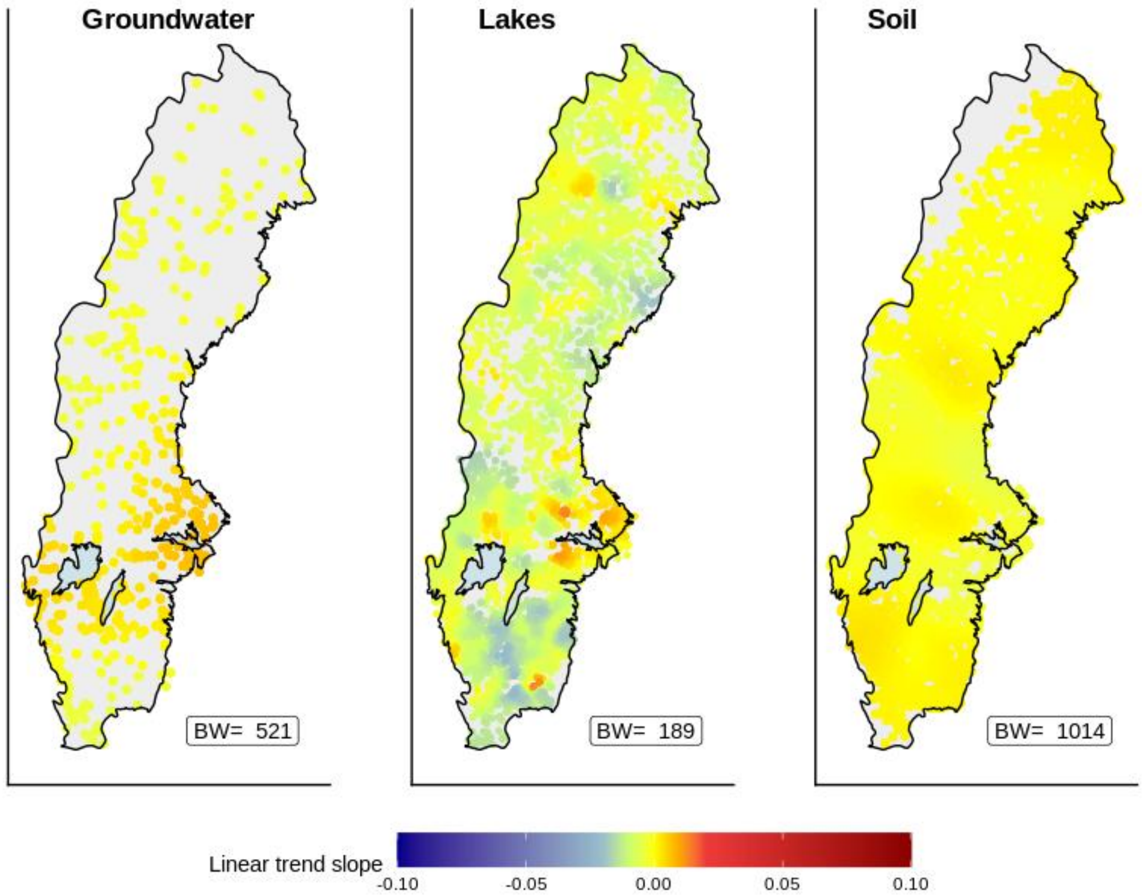


Figure 2: Geographically weighted trends for pH after adjustment for trends in calcium. Left: Groundwater (2007-2024), Middle: Lakes (2008-2024), Right: Soil (2003-2023)

Investigating the relationship between pH and Ca over the country we found the relationship to be mostly positive, i.e. higher values of Ca arose with higher values of pH (Figure 3), but there were some areas that deviated. For groundwater and lakes this included the Stockholm area, where the relationship between Ca and pH was negative. However, the region where these negative relationships were observed did not overlap completely for the two media. For soils the relationship between these two variables was not as strong. The highest regression coefficients were observed in the North.

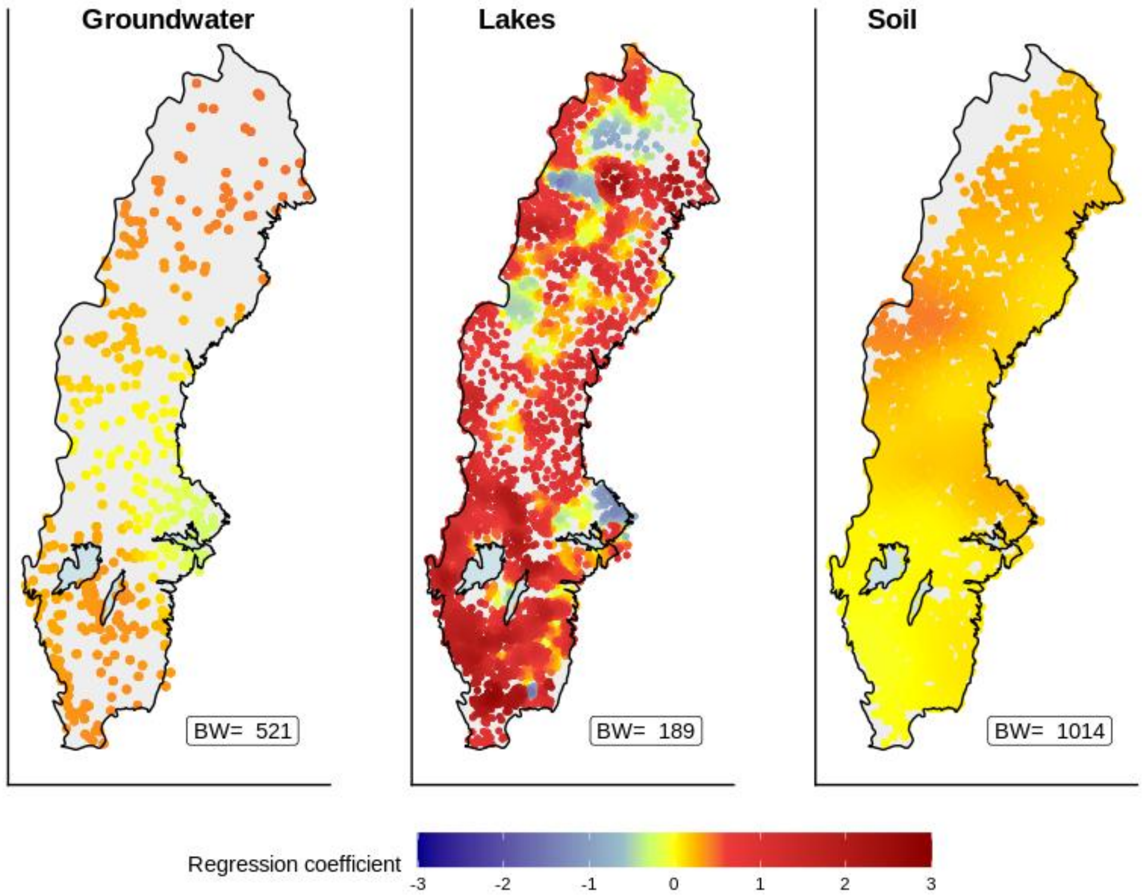


Figure 3: Geographically weighted coefficient for the relationship between pH and calcium. Left: Groundwater (2007-2024), Middle: Lakes (2008-2024), Right: Soil (2003-2023)

Multiscale geographically weighted regression models for trend in pH using time and calcium as explanatory variable

If the model additionally allows the use of different bandwidth for the estimation of trend and the estimation of the relationship between pH and Ca the results become more local and more difficult to interpret (Figure 4). For example, for groundwater no trends were observed anymore, while the relationship between pH and Ca was locally strong with varying direction (Figure 5). For lakes the results were similar, but some more general features from earlier models were still retained. For soil, again the strongest relationship between Ca and pH were positive and observed in the North, and most increasing trends were again observed in the North, but with much more local variation.

The bandwidths determined for these models were generally low for the relationships between Ca and pH, indicating very local estimates. They ranged between 10 and 16, which means that approximately 5 sites were used in each of the estimations.

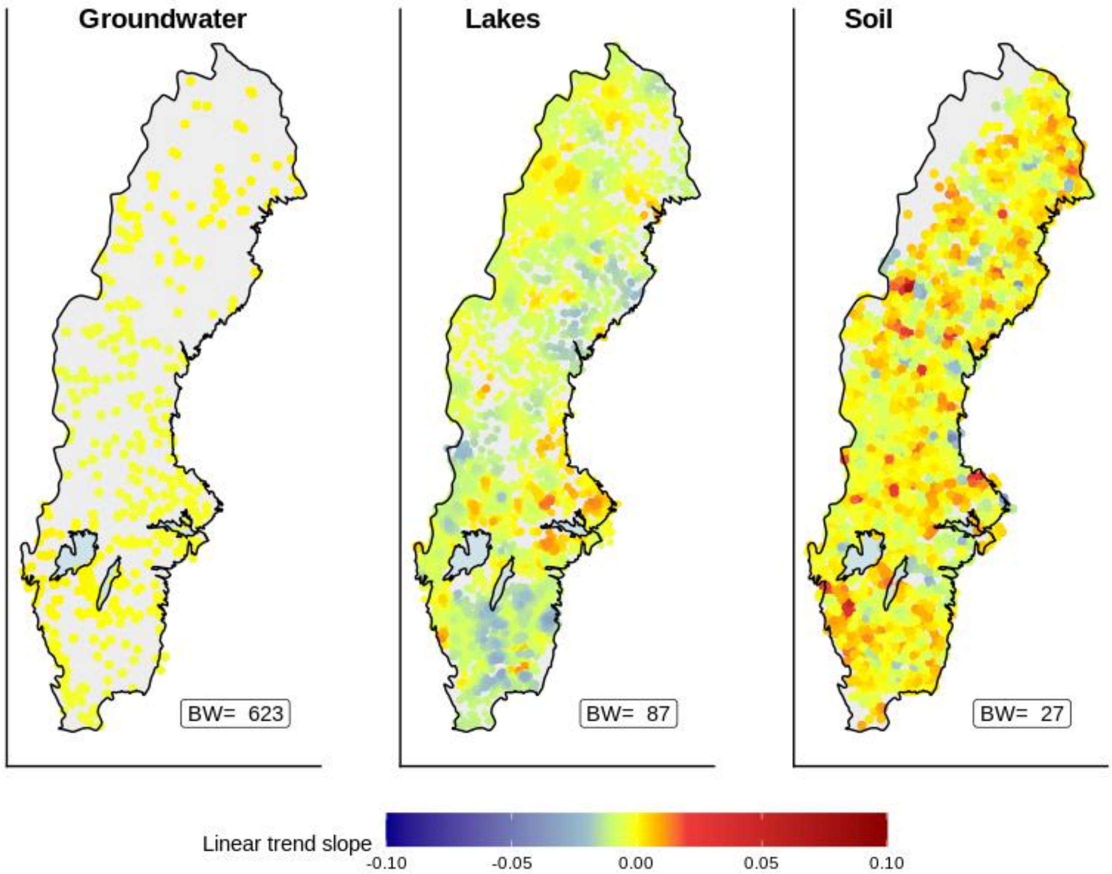


Figure 4: Geographically weighted trends for pH after adjustment for trends in calcium. Left: Groundwater (2007-2024), Middle: Lakes (2008-2024), Right: Soil (2003-2023)

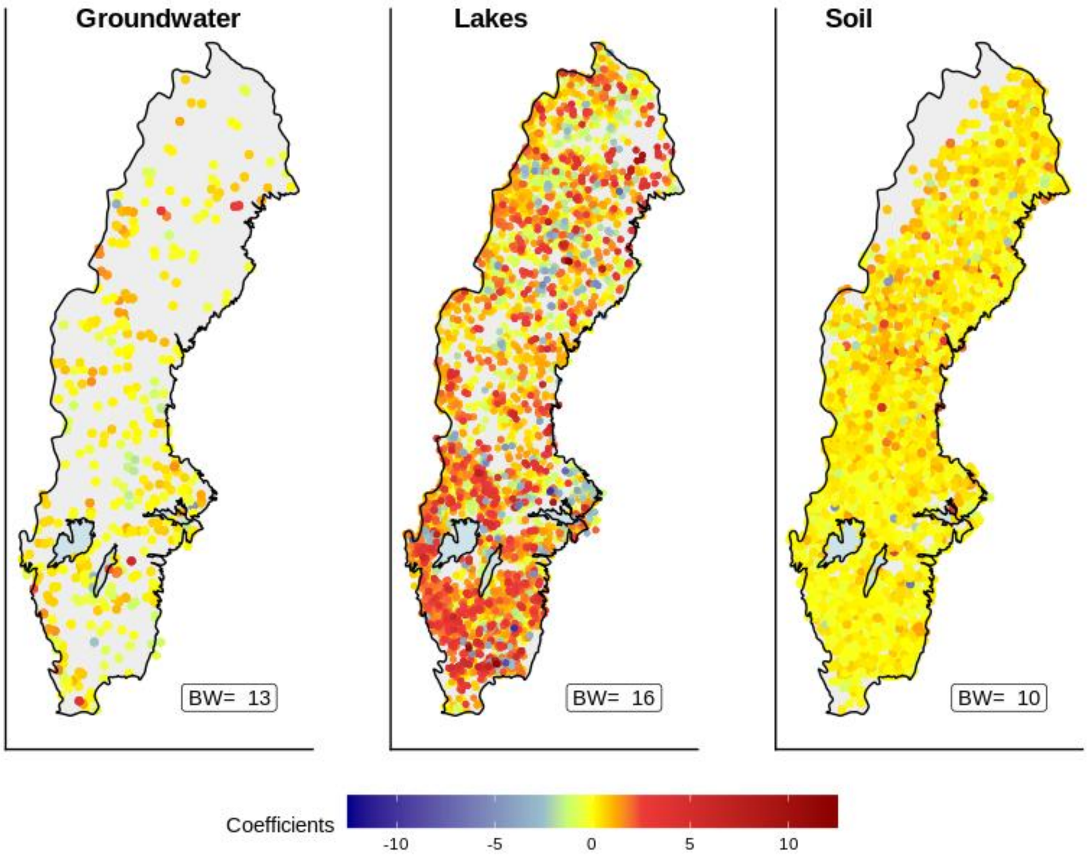


Figure 5: Geographically weighted coefficient for the relationship between pH and calcium. Left: Groundwater (2007-2024), Middle: Lakes (2008-2024), Right: Soil (2003-2023)

Quantitative model comparisons

GWR should be seen as a method focused on visualization of data. There are some goodness-of-fit measures that can be computed (Table 1), but which need to be interpreted with care. For example, we receive an overall R^2 value, i.e. a quantification of how much variation in the response variable is explained by the model. This measure is generally quite low for the basic model and the basic model including Ca. This does, however, not necessarily indicate that the models are not good, as a situation where no or weak trends are present in most of the country would hold this measure down, even if substantial and interesting trends are detected in certain areas. Maximum local R^2 values could be used to illustrate local success of the model. These reach 0.7 – 0.84 for the basic models with Ca in lakes, and 0.37 for groundwater. The ability to explain variation in the soil dataset is generally low for all models and only reaches 0.27 for the basic model with Ca, which is observed for sites in the North (Figure 6). For groundwater the model shows the highest R^2 values in the North and in the far South, while for lakes the highest local R^2 values are noted for the South-West and some areas in the North. For multiscale GWR the local R^2 would be expected to be very high in certain areas, but this measure is not yet implemented (Fotheringham et al., 2017).

For application purposes we can also compare runtimes of the model, which is very dependent on the number of observations in the dataset. For the groundwater dataset with about 400 stations runtime is no issue, while computation for lakes and soils (~4600 lakes and ~3200 soil sites) takes more time, especially when a multiscale model is used. However, runtime was not the major issue for this model, but the memory use that is so substantial that it during much of the computation time of multiscale model for lakes and soil was not possible to use the computer for other tasks.

Table 1: Adjusted R^2 for the model, maximum local R^2 and total runtime.

		Adjusted R^2	Max local R^2	Run time
Basic	Ground water	0.001	0.011	~ 1 sec
	Lakes	0.056	0.182	~2 mins
	Soil	0.000	0.001	~4 secs
Basic with Ca	Ground water	0.160	0.702	~1 sec
	Lakes	0.258	0.840	~3 mins
	Soil	0.102	0.281	~4 secs
Multiscale with Ca	Ground water	0.489		~2 mins
	Lakes	0.425		~ 3 hours
	Soil	0.594		~20 mins

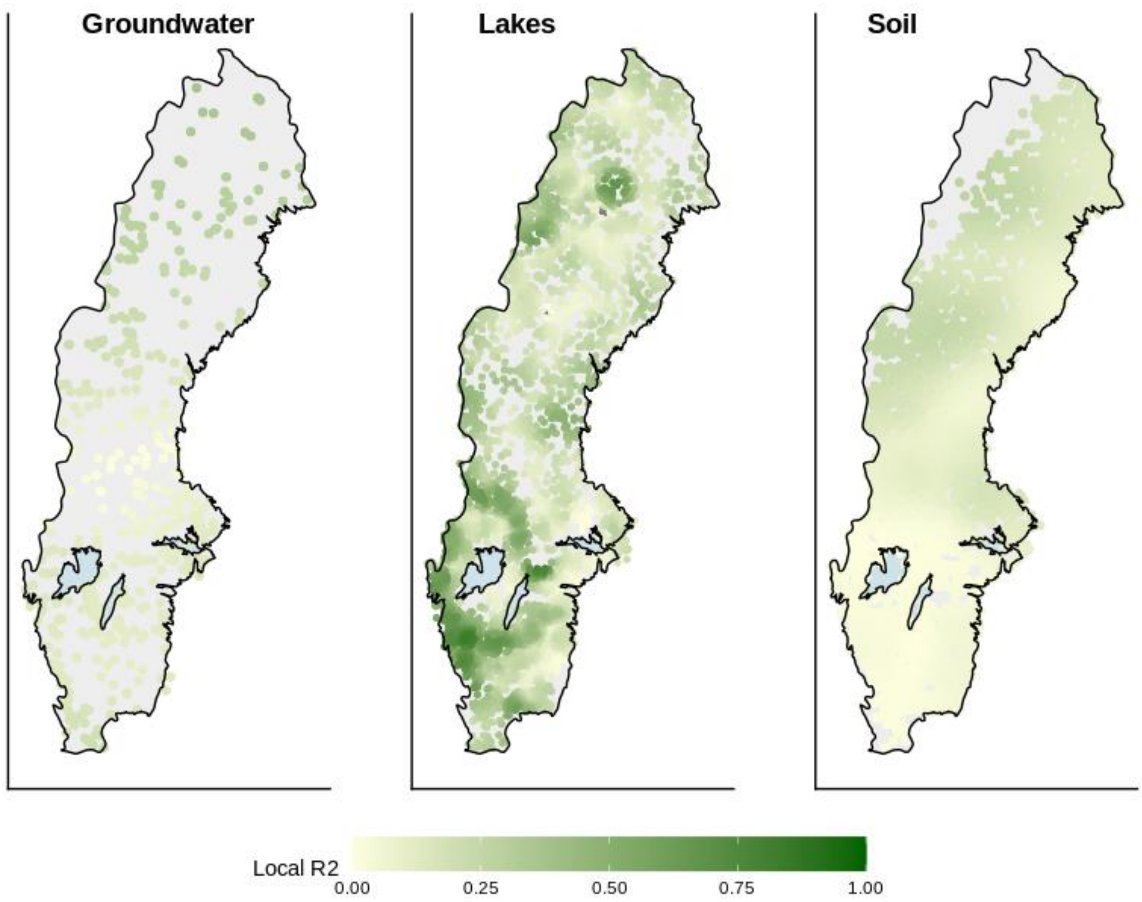


Figure 6: Local R2 values of the basic geographically weighted coefficient including Ca. Left: Groundwater (2007-2024), Middle: Lakes (2008-2024), Right: Soil (2003-2023)

Discussion

The goal of this report was to investigate the potential of multiscale geographically weighted regression models to evaluate the connection to one or several explanatory variables. We encountered two major problems:

(i) Multiscale GWR models are computationally heavy. We have earlier tried to reduce necessary run time and memory demands of the GWR functions (Höglund and von Brömssen, 2025). Still, the time necessary for the model to run with lakes data (~4600 lakes) or soil data (~3200 sites) using only one explanatory variable in addition to time, is between 20 minutes and three hours, which makes it impractical for researchers. Moreover, memory use is still excessive, which makes it difficult to run such models on a work computer.

(ii) The estimated bandwidth for the explanatory variable, calcium, was generally very small, indicating that the relationship between pH and calcium is fitted locally. The relationship also varies throughout the country with both positive and negative coefficients. This is most likely an overfitting effect. The model we use is very complex and can adjust well to data. While the best model is determined by cross-validation, which should avoid over-fitting to some extent, we also need to consider that the underlying data is very sparse and might not provide enough information for such detailed modelling.

With these results the overall recommendation is to not use multiscale GWR models for the present data types, i.e. environmental monitoring data collected in “omdrev” as the overall data support must be considered too low for trend analysis if observations are only made once each six or ten years. Instead, it could be interesting to test the multiscale model on datasets that are more frequently measured in time but still have a dense enough spatial resolution.

The basic model should be explored further for different selections of response and explanatory variables in cases where we can assume that these variable(s) and the time-trend work on similar scales, e.g. if all of those are either global, regional or local. The basic models are also much less computationally heavy, which allows comparison of different models, e.g. by including different explanatory variables.

We found that the relationship between Ca and pH was strongest in lakes, where usually high pH levels were accompanied by high values of Ca. The relationships were weaker and more diverse for groundwater sites, which positive relationships mainly in the North and South. For soil the relationships were even less pronounced. However, for this model the smoothing was also more substantial, which would lead to smaller estimated regression coefficients.

The least specific results were obtained for soils, which is not surprising as there typically is more small-scale variation between different soil samples. Improved models could be defined by selecting more homogeneous subsets from this dataset to not mix different types of sources of variation. For soil, with a revisit time of ten years, the period to be included in the analysis is longer, starting in 2003, while for groundwater the first year included is 2007 and for lakes 2008.

In the models applied it is assumed that the temporal trends for pH are linear over this monitoring period, i.e. during the last 16-20 years. This assumption might also be a reason why only weak trends are detected. While the analysis could be run on 12 years of data for lakes and groundwater to catch more current trends, for the soil data to obtain two observations per site an observation period of 20 years is needed and for comparison the longer series for groundwater and lakes are also retained.

Funding

This study was funded by SLUs grand for environmental assessment 2025 within the assessment area of acidification. Work on refinement of multiscale GWR computational aspects was funded through the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (Formas) (Grant No. 2022–00942).

Code and Data Availability

The code used for computation can be found at https://github.com/claudiavonbromssen/Multiscale-GWR-for-trends-in-groundwater-lakes-and-soils. Datasets can be obtained from the authors.

References

Brunsdon, C., Fotheringham, S., Charlton, M., 1998. Geographically Weighted Regression. J. R. Stat. Soc. Ser. Stat. 47, 431–443. https://doi.org/10.1111/1467-9884.00145

Erlandsson, M., Cory, N., Fölster, J., Köhler, S., Laudon, H., Weyhenmeyer, G.A., Bishop, K., 2011. Increasing Dissolved Organic Carbon Redefines the Extent of Surface Water Acidification and Helps Resolve a Classic Controversy. BioScience 61, 614–618. https://doi.org/10.1525/bio.2011.61.8.7

Fölster, J., Hallstan, S., Johnson, R., 2014a. Utvärdering av de nationella miljöövervakningsprogrammen av sjöar (Rapport No. 2014:3). Uppsala.

Fölster, J., Johnson, R.K., Futter, M.N., Wilander, A., 2014b. The Swedish monitoring of surface waters: 50 years of adaptive monitoring. AMBIO 43, 3–18. https://doi.org/10.1007/s13280-014-0558-z

Fotheringham, A.S., Yang, W., Kang, W., 2017. Multiscale Geographically Weighted Regression (MGWR). Ann. Am. Assoc. Geogr. 107, 1247–1265. https://doi.org/10.1080/24694452.2017.1352480

Garmo, Ø.A., Skjelkvåle, B.L., de Wit, H.A., Colombo, L., Curtis, C., Fölster, J., Hoffmann, A., Hruška, J., Høgåsen, T., Jeffries, D.S., Keller, W.B., Krám, P., Majer, V., Monteith, D.T., Paterson, A.M., Rogora, M., Rzychon, D., Steingruber, S., Stoddard, J.L., Vuorenmaa, J., Worsztynowicz, A., 2014. Trends in Surface Water Chemistry in Acidified Areas in Europe and North America from 1990 to 2008. Water. Air. Soil Pollut. 225, 1880. https://doi.org/10.1007/s11270-014-1880-6

Gollini, I., Lu, B., Charlton, M., Brunsdon, C., Harris, P., 2015. GWmodel: An R Package for Exploring Spatial Heterogeneity Using Geographically Weighted Models. J. Stat. Softw. 63, 1–50. https://doi.org/10.18637/jss.v063.i17

Göransson, E., Johnson, R.K., Wilander, A., 2004. Representativity of a Mid-lake Surface Water Chemistry Sample. Environ. Monit. Assess. 95, 221–238. https://doi.org/10.1023/B:EMAS.0000029905.07072.b0

Grennfelt, P., Engleryd, A., Forsius, M., Hov, Ø., Rodhe, H., Cowling, E., 2020. Acid rain and air pollution: 50 years of progress in environmental science and policy. Ambio 49, 849–864. https://doi.org/10.1007/s13280-019-01244-4

Höglund, H., von Brömssen, C., 2025. Reducing the running time for gwr.multiscale in GWmodel (No. 2025:130). Department of Energy and Technology, Swedish University of Agricultural Sciences.

Minella, M., De Laurentiis, E., Buhvestova, O., Haldna, M., Kangur, K., Maurino, V., Minero, C., Vione, D., 2013. Modelling lake-water photochemistry: Three-decade assessment of the steady-state concentration of photoreactive transients in the surface water of polymictic Lake Peipsi

(Estonia/Russia). Chemosphere 90, 2589–2596. https://doi.org/10.1016/j.chemosphere.2012.10.103
R Core Team, 2025. R: A Language and Environment for Statistical Computing.
von Brömssen, C., Betnér, S., Fölster, J., Eklöf, K., 2021. A toolbox for visualizing trends in large-scale environmental data. Environ. Model. Softw. 136, 104949. https://doi.org/10.1016/j.envsoft.2020.104949
von Brömssen, C., Fölster, J., Eklöf, K., 2025. Thematically weighted regression models for identification of important drivers of environmental trends in lake survey data. Environ. Monit. Assess. 197, 1212. https://doi.org/10.1007/s10661-025-14611-4
von Brömssen, C., Fölster, J., Eklöf, K., 2023. Temporal trend evaluation in monitoring programs with high spatial resolution and low temporal resolution using geographically weighted regression models. Environ. Monit. Assess. 195, 547. https://doi.org/10.1007/s10661-023-11172-2
Vuorenmaa, J., Augustaitis, A., Beudert, B., Bochenek, W., Clarke, N., de Wit, H.A., Dirnböck, T., Frey, J., Hakola, H., Kleemola, S., Kobler, J., Krám, P., Lindroos, A.-J., Lundin, L., Löfgren, S., Marchetto, A., Pecka, T., Schulte-Bisping, H., Skotak, K., Srybny, A., Szpikowski, J., Ukonmaanaho, L., Váňa, M., Åkerblom, S., Forsius, M., 2018. Long-term changes (1990–2015) in the atmospheric deposition and runoff water chemistry of sulphate, inorganic nitrogen and acidity for forested catchments in Europe in relation to changes in emissions and hydrometeorological conditions. Sci. Total Environ. 625, 1129–1145. https://doi.org/10.1016/j.scitotenv.2017.12.245

Supplementary

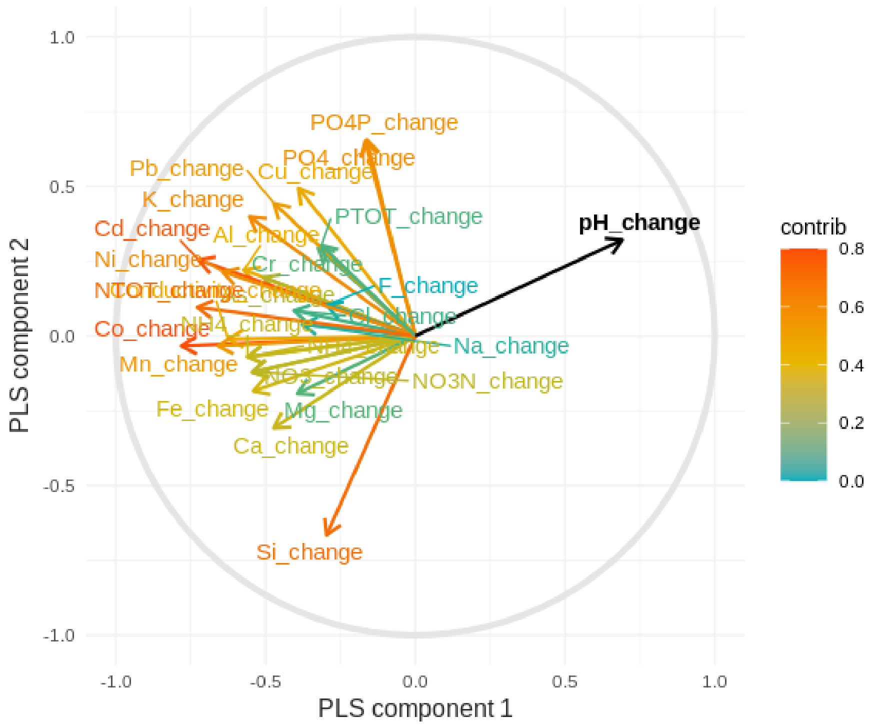


Figure S.1: PLS components illustrating relationships between pH change and other chemical variables in groundwater.

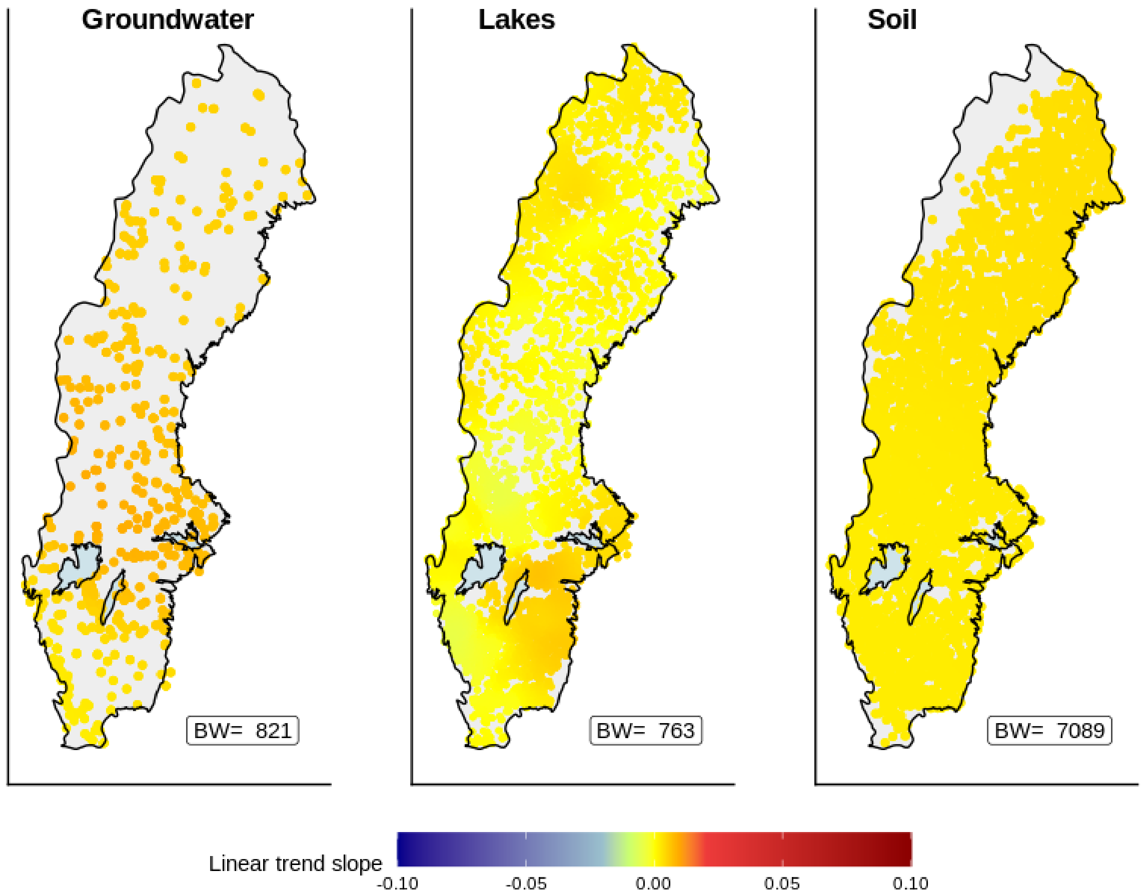


Figure S2: Geographically weighted trends for Ca. Left: Groundwater (2007-2024), Middle: Lakes (2008-2024), Right: Soil (2003-2023)