



Decomposition of wastewater sources in England



Barnaby Dobson^{1,2}✉, Robin Maes-Prior² & Christopher R. Jackson³

The flow arriving at a wastewater treatment plant mixes municipal sewage, rainfall, soil moisture and groundwater, yet is routinely interpreted as if its composition were fixed. Despite the uncertainty this introduces to any case where the signal is used, from combined sewer overflow management to wastewater-based epidemiology, source-resolved composition is relatively underexplored. Here we introduce WWCompose, an empirical-conceptual model that recovers the dynamic contributions of rainfall, soil moisture and groundwater to daily influent, and apply it to approximately 2900 treatment plants across England using open national data. Predicted flows match observations out of sample (median held-out Kling–Gupta efficiency 0.64; 60% of plants above 0.6), with fitted parameters organising around hydrogeology, permeability and urbanisation. Non-municipal water is widespread: municipal flow averages 61% nationally, and at individual works may vary from 100% in dry weather to <5% during the wet conditions that often dominate periods of interest.

Influent arriving at a Wastewater Treatment Plant (WWTP) constitutes water from multiple sources and pathways. It is composed of household and trade wastewater, direct storm water runoff and drainage, rainfall-derived inflow and infiltration¹, and groundwater². The composition of influent matters because the same observed flow can create different interpretations for flooding³, combined sewer overflow spills⁴, diffuse urban pollution⁵, receiving-water exposure⁶, wastewater-based epidemiology⁷, and the design or operation of treatment technology⁸. For example, the consequences of a sewer spill depend strongly on whether the discharged flow is primarily rainwater or municipal effluent⁵, while interpretation of SARS-CoV-2 wastewater samples depends on the number of people contributing to the influent⁹. Both cited examples implicitly assumed a static influent composition, introducing significant uncertainty given its known dynamic variability¹⁰. Rainfall-derived inflow and infiltration, and groundwater sources are established as major influent contributors^{11,12}, observed to compose up to 70% of influent at some WWTPs^{11,13}. However, they are difficult to estimate owing to multiple data, scale, and conceptual limitations¹⁴. The mechanism for groundwater ingress is that the water table rises above some pipes in a sewer network and enters through cracks or gaps¹⁵; this is categorised as phreatic zone sewer infiltration¹⁴. Although this process is conceptually straightforward, variability in pipe depths and groundwater depth make quantifying phreatic zone sewer infiltration incredibly challenging from a data perspective¹⁶. Rainfall-derived inflow and infiltration is referred to as vadose zone sewer infiltration¹⁴, which captures a diverse set of potential processes in the near-surface such as preferential flow paths¹⁷ and connectivity with hydrological features such as rivers^{18,19}. Although characterising each mode completely requires comprehensive

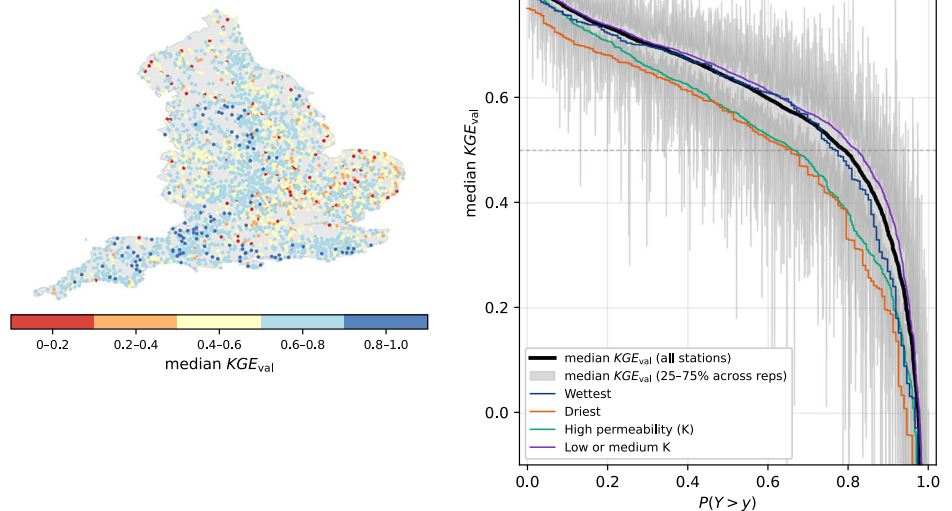
data and modelling efforts¹⁶, dynamic variability in phreatic zone infiltration can generally be attributed to variability in groundwater level²⁰ and variability in the vadose zone to soil moisture²¹.

A combination of physical modelling²², statistical data analysis and machine learning^{23,24}, and conceptually based empirical methods²⁵ has been demonstrated to achieve good predictive performance of WWTP influent in literature. These studies are typically focused on few WWTPs further highlighting the data limitations in sewer infiltration. The only studies that investigate infiltration on more than 10 WWTPs^{26–28}, do not make predictions about WWTP influent and so cannot be validated. Furthermore, none of these larger scale studies examine the implications of influent composition as part of their modelling approaches, focussing instead solely on flow prediction.

Quantifying influent composition currently requires intensive field surveys involving tracers²⁹, which do not scale to the hundreds or thousands of WWTPs at national scale. The clearest demands for composition estimates are inherently large-scale, for example, in the UK, successive governments have committed £60 billion to the alleviation of sewer spills³⁰. However, appropriate interventions differ fundamentally depending on composition; pipe-lining might reduce groundwater-driven spills but it would not reduce precipitation-driven spills. Similar arguments apply to wastewater-based epidemiology, where the human-derived fraction of a sample governs its interpretation, and to the representation of wastewater in hydrological models. These applications motivate (i) a simulation tool that does not depend on explicit sewer-network topology or site-specific campaigns, (ii) its application at national scale, and (iii) a characterisation of the

¹Trinity College Dublin, Dublin, Ireland. ²Imperial College London, London, UK. ³British Geological Survey, Keyworth, UK. ✉e-mail: bdobson@tcd.ie

Fig. 1 | National model performance across England WWTPs. Left: median held-out Kling-Gupta efficiency (KGE) for each plant, over the ten cross-validation blocks. Right: exceedance probability (x -axis) for a given KGE value (y -axis), different coloured lines show different groups, while the grey band shows the 25th–75th percentile across cross-validation repetitions.



catchment factors that control where each infiltration mechanism dominates, so that behaviour can be anticipated at unmodelled locations.

In this study, we identify that the open data environment in England currently supports a nationwide assessment and validation of WWTP influent composition in 2,900 locations. We build on existing conceptually based empirical methods²⁵ to test the hypothesis that daily influent flow contains recoverable signatures from dynamic data on rainwater, soil moisture, and groundwater level. The outputs of such an approach facilitate quantifying dynamic wastewater composition with applications ranging from sewer spills to wastewater epidemiology.

Results

National model performance

We validate our approach by comparing predicted and observed daily flow at WWTPs across England (Fig. 1), reporting for each WWTP the median KGE over the ten held-out cross-validation blocks (Methods, Calibration and validation). The national median is 0.64 (inter-station IQR 0.53 – 0.72), 60% of works exceed 0.6 and 87% exceed 0.4. For comparison with studies reporting traditional metrics, the national median R^2 between simulated and observed flows is 0.55; station rankings under R^2 and KGE agree almost exactly (Spearman $\rho = 0.99$), so the choice of metric does not affect our conclusions. The spread over cross-validation repetitions is narrow relative to the national performance gradient, indicating that the fitted decompositions are not dominated by arbitrary split-to-split variability. The lowest-performing works are interpretable failures rather than random outliers: visual inspection commonly shows operational signatures in the effluent record; primarily abrupt changes in mean flow or variance. The main medium-performance band $KGE \approx 0.4$ – 0.6 constitutes 28% of WWTPs and is concentrated in the east of England, consisting of smaller plants with more variable operational regimes and drier catchments. Held-out performance increases with record length (Spearman $\rho = 0.33$), works size ($\rho = 0.34$, taking the fitted municipal constant as a size proxy) and catchment wetness ($\rho = 0.25$); these three factors are only weakly related to one another ($|\rho| \leq 0.17$). Drier climates and more permeable geology are typically harder to fit, with the driest 5% of catchments (those with average catchment daily rainfall below the 5th percentile across all stations, compared against all remaining stations) having a median held-out $KGE = 0.57$ and higher permeability catchments (compared against low and medium permeability combined) $KGE = 0.58$, see Fig. 1 and the Parameter and composition variability section for further details.

Spatial structure in fitted parameters

The fitted parameters exhibit distinct spatial organisation (Fig. 2) varying from parameter to parameter. For example, the constant scaling parameter for groundwater contributions C_{GW} broadly traces major aquifers (Spearman $\rho = 0.16$ against the presence of a highly productive aquifer³¹, $p < 10^{-17}$), the constant scaling parameter for municipal contributions C_{Mun} increases around urban areas ($\rho = 0.89$ against catchment impervious area), and the exponent applied to precipitation contributions E_P increases in the wettest regions ($\rho = 0.20$ against mean catchment rainfall, $p < 10^{-25}$). If wastewater catchments fell into a small number of archetypes – recurring joint configurations of municipal, groundwater, soil-moisture and rainfall behaviour – the fitted parameters would exhibit strong low-dimensional structure, with a few principal components explaining most inter-station variance and stations separating into distinct clusters. Neither is observed. Although some common structure exists across fitted parameters, it is weak: the top four components of a principal component analysis explain 47% of the variance across the 17 fitted parameters, and the top eight explain 73%. K-means clustering of the same standardised parameters identifies only two separable groups, corresponding to works without a usable borehole and to works whose soil parameters sit at extreme SoilGrids-derived bounds; both groupings are imposed by data availability rather than fitted behaviour. We therefore find no support for distinct wastewater catchment archetypes: catchments occupy a continuum organised by individual controls (aquifer presence, urbanisation, wetness). Weak structure may also partly reflect fitting noise across stations (see Discussion). The upper precipitation gate threshold $T_{p, gate}^{hi}$, the cap on rainfall excess T_p^{ub} , field capacity θ_f , saturated water content θ_s , and wetting-front suction head ψ are omitted from Fig. 2 because they show limited variability or no distinctive national spatial pattern.

National composition

The fitted models derive the composition fractions for a given WWTP on a given day. These daily compositions are averaged over the entire data period, 2011 – 2023, and shown in Fig. 3. At a national scale, municipal is the largest average contribution (61%; inter-station IQR 54–71%), followed by soil moisture (20%; 13–26%), groundwater (12%; 0.1–18%), and precipitation (7%; 4–9%). Despite its prevalence, only 0.1% of catchments are >95% municipal, highlighting the rarity of a textbook fully separate sewer system. Sensible spatial groupings are evident – we see higher precipitation composition in the wetter west, lower soil moisture composition upon chalk

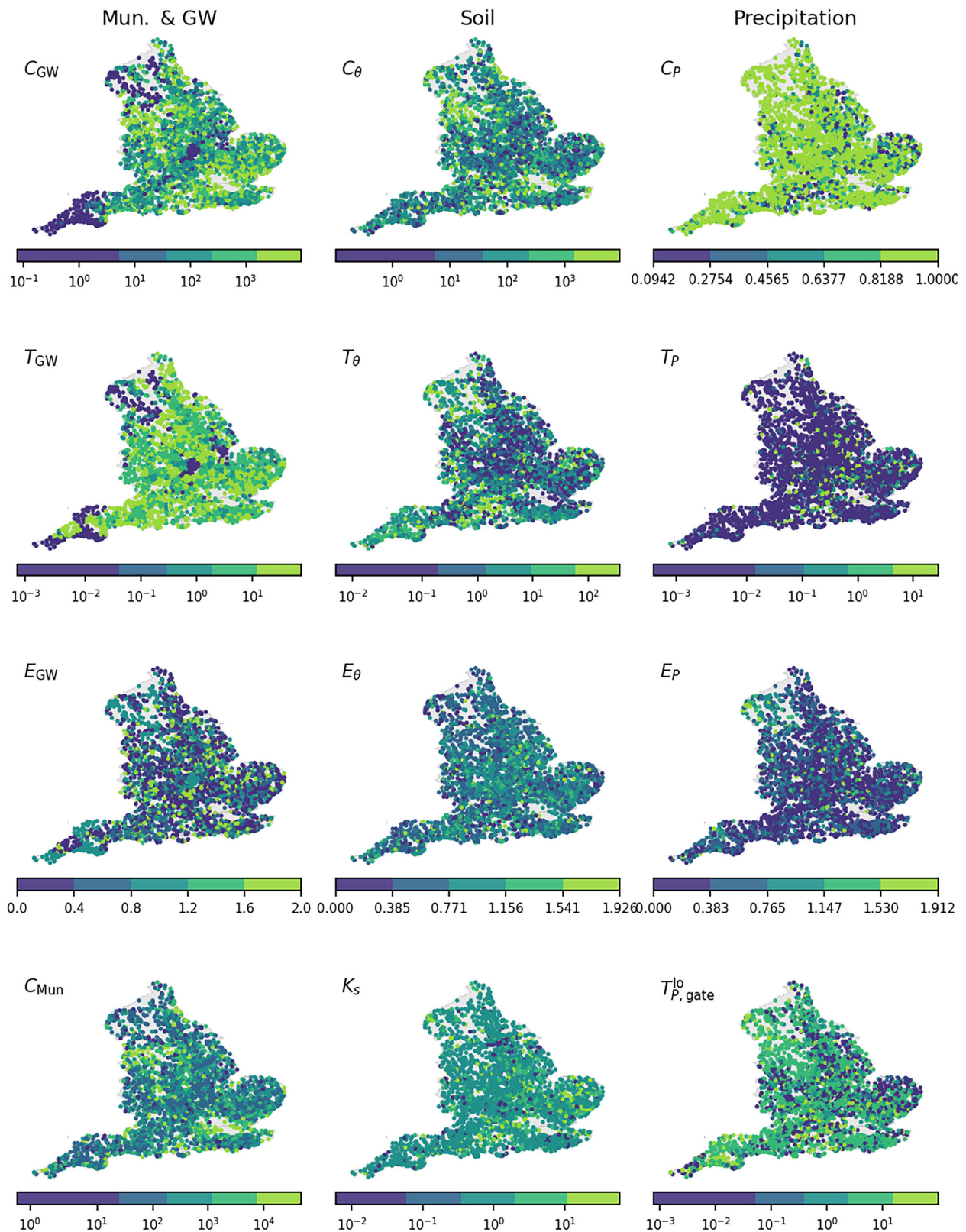


Fig. 2 | Spatial structure in fitted model parameters. Per-WWTP average parameter values. GW stands for groundwater, θ for soil moisture, P for precipitation, Mun for municipal. C refers to a linear factor applied to that model component, T a

threshold above which a contribution must rise to contribute to influent, and E an exponent applied to contributions above T . K_s refers to saturated conductivity. Further details are provided in Methods.

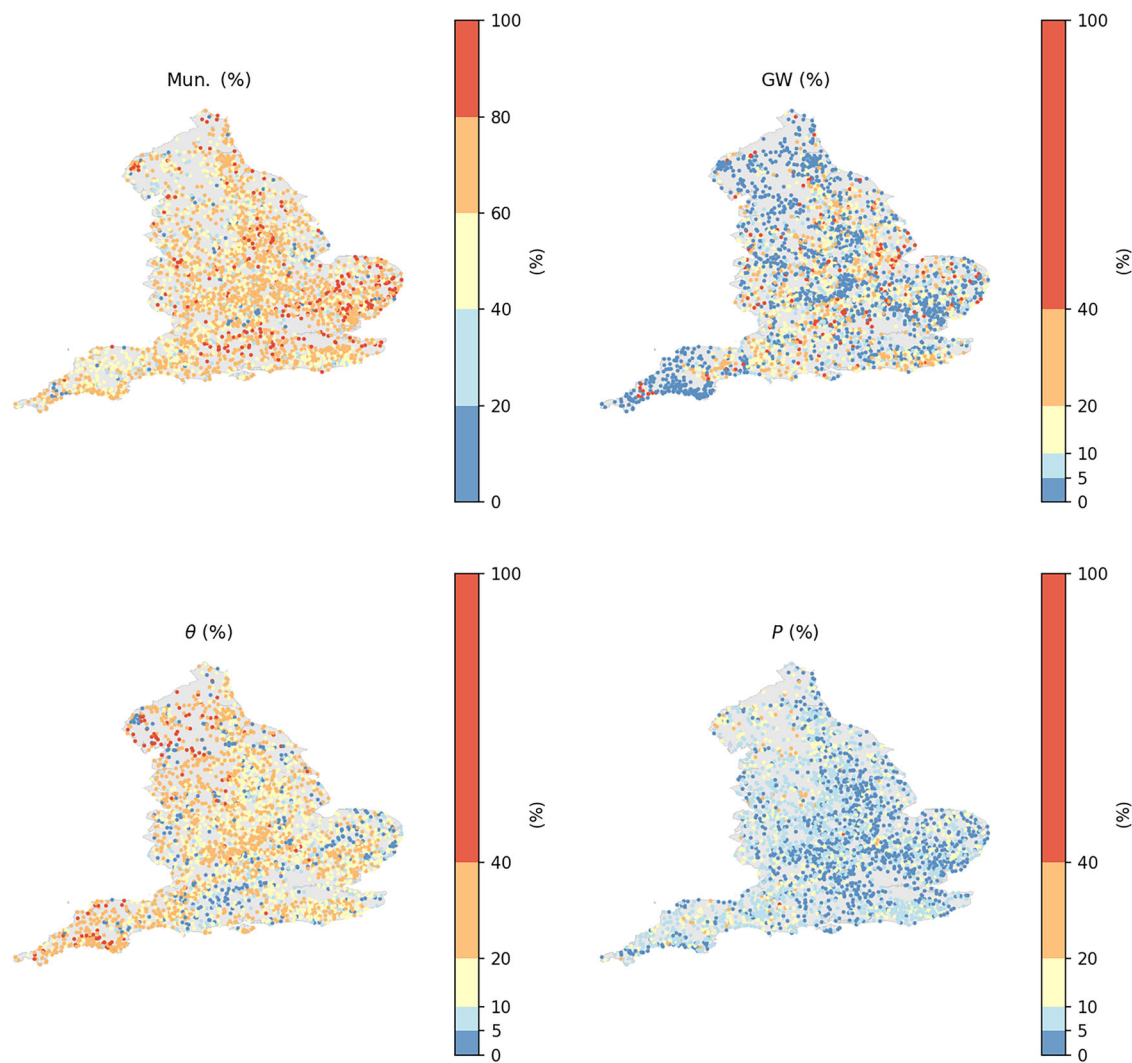


Fig. 3 | National wastewater composition. Per-plant average composition. *Mun.* stands for municipal, *GW* for groundwater, θ for soil moisture, and *P* for precipitation.

(Spearman $\rho = -0.23$ against chalk rock units³¹, $p < 10^{-36}$), and higher groundwater composition above major aquifers (quantified in the Parameter and composition variability section)—but compositions within these categories are still highly variable. The groundwater range additionally reflects the 10% of works without a usable borehole (Methods, Data sources and alignment). Municipal-dominated works in the drier east coincide with the medium-performance band of Fig. 1, consistent with the weaker climate-driven signal available for fitting there.

The decompositions are robust to the period used for fitting: across the ten cross-validation fits, the national-median spread (interquartile range) of a work's average composition is 2.5 percentage points for the municipal fraction, 1.7 for groundwater, 1.6 for soil moisture and 0.8 for precipitation. The minority of works with spreads above 10 percentage points coincide with the lower-performance sites of Fig. 1 (median held-out KGE 0.57 versus 0.65 elsewhere). Supplemental Fig. 1 shows these bounds at the three case-study works: the decomposition is well constrained at Westwood Brook and Oakley, whereas at Chesham we see how the groundwater component can, in some repetitions, consume the entire municipal component.

Parameter and composition variability

Figures 1 and 3 indicate the importance of aridity and permeability towards model behaviour. We investigate these categories in greater detail with Fig. 4, which shows, for each fitted parameter, average composition fraction, and held-out performance value, the signed two-sample

Kolmogorov–Smirnov statistic between each stratum (high permeability; driest 5%; wettest 5%) and all remaining stations (Methods, Analysis of fitted models); values beyond the dashed lines differ significantly at the 5% level. No normalisation is applied before the tests; the Kolmogorov–Smirnov statistic depends only on the ordering of values within each sample. For absolute-magnitude parameters, principally C_{Mun} , group differences therefore partly reflect systematic differences in works scale between strata, and we interpret these rows accordingly. The results support the interpretation of model behaviour, below.

Greater permeability is associated with reduced soil-moisture and rainfall composition, $\theta(\%)$ and $P(\%)$, consistent with greater opportunity for surface-water dissipation before reaching the sewer network. Reduced contributions from these sources correspond to increased composition proportions of groundwater and municipal, $GW(\%)$ and $Mun(\%)$. Municipal composition increases by more than groundwater composition. We interpret this as reflecting sewer depth: sewers are typically laid at depths of 0.5–5 m³² that are above the water table across much of England's permeable geology, whereas nearly all works have a significant municipal component. Consistent with this, at the high-permeability exemplar Chesham the assigned borehole records water tables at 38–42 m depth, implying that even where chalk infiltration is confirmed by field samples it is localised rather than network-wide. Practically, this suggests that infiltration reduction in permeable catchments should target specific below-water-table assets rather than network-wide rehabilitation. We see that some

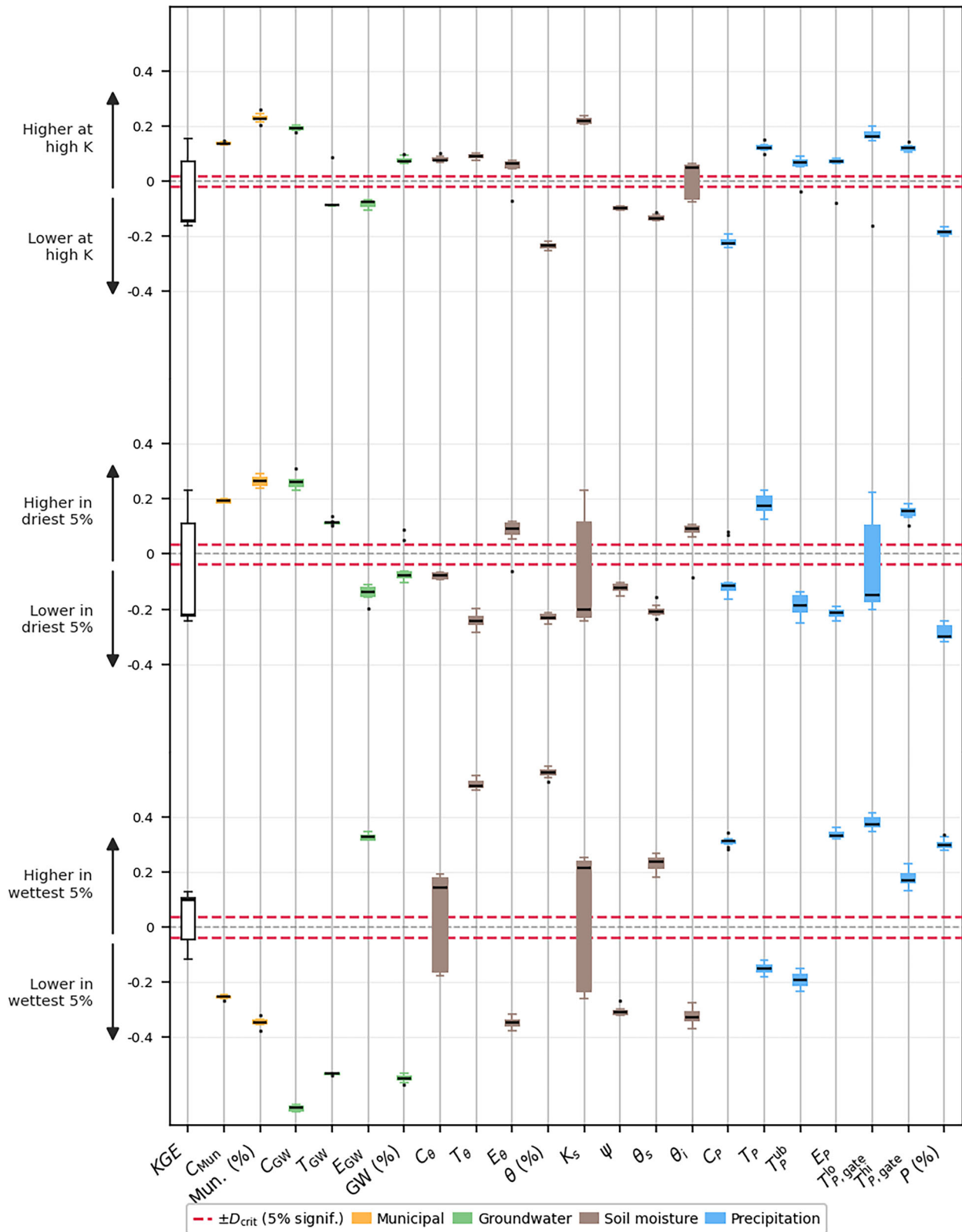


Fig. 4 | Signed Kolmogorov-Smirnov statistics comparing fitted parameters, average compositions, and held-out performance between strata and all remaining stations. Top row: high vs low/medium permeability; middle row: driest

5%; bottom row: wettest 5%. Boxes show the distribution over the ten cross-validation replicates; dashed lines mark the 5% significance threshold.

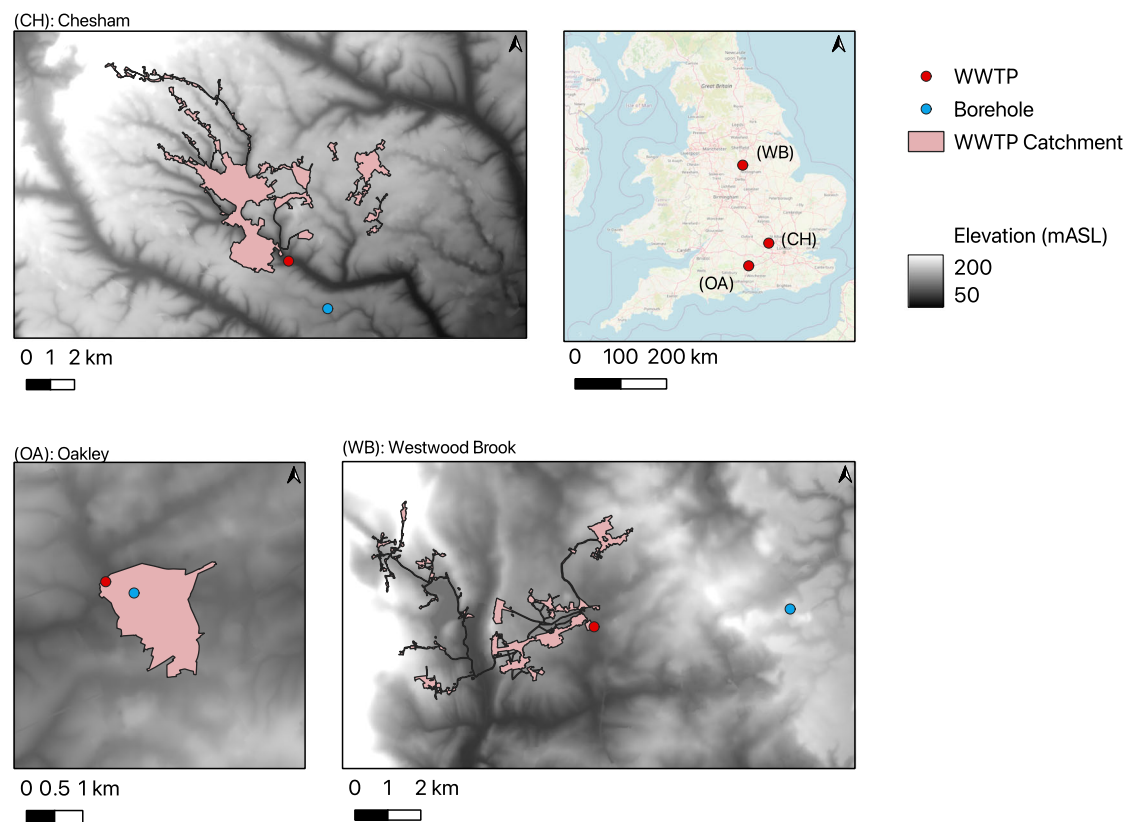


Fig. 5 | Map of exemplar case study locations. Selected WWTPs span contrasting regions. Elevation data is from Terrain Tiles, accessed on 2026-06-01 from <https://registry.opendata.aws/terrain-tiles>. The inset basemap is © OpenStreetMap contributors, available under the Open Database Licence (<https://www.openstreetmap.org/copyright>).

parameters contributing to soil moisture (the constant scaling parameter for soil moisture contributions C_θ , the exponent applied to soil moisture contributions E_θ , saturated hydraulic conductivity K_s) do increase, but also that the soil moisture contribution threshold T_θ increases, consistent with weaker overall soil-moisture contributions (decrease in $\theta(\%)$) that occur in a more peaked manner.

WWTPs in drier areas show higher municipal composition and reduced composition from all other sources. Reductions in T_θ and the lower precipitation gate threshold $T_{p, \text{gate}}^{\text{lo}}$, and increases in $T_{p, \text{gate}}^{\text{hi}}$ and E_θ are consistent with more variable but weaker (decreased C_θ and the constant scaling parameter for precipitation contributions C_p) contributions from direct and indirect soil moisture. In contrast, reductions in the exponent applied to groundwater contributions E_{GW} and E_p and increases in the groundwater contribution threshold T_{GW} and the precipitation contribution threshold T_p are consistent with decreased variability in contributions from groundwater and precipitation. Wetter WWTPs generally show the reverse associations, with higher precipitation and direct soil-moisture contributions and reduced municipal, groundwater and indirect soil-moisture contributions.

Case studies

Three exemplar sites are demonstrated with timeseries to illustrate visually how the observation data is incorporated into WWCompose and the importance of dynamic variability when interpreting decomposition results, shown in Fig. 5. The sites exhibit a variety of behaviours and cover different climate and hydro-geological conditions. Westwood Brook, a high-performing low-permeability site, demonstrates a balanced composition varying strongly in time – the situation in which decomposition is most consequential. Chesham, a high-permeability chalk site with field-confirmed infiltration, isolates groundwater-level-driven behaviour and the limits of single-borehole forcing. Oakley combines ephemeral groundwater overtopping with the difficulty of fully capturing precipitation-driven

response. Together they span the permeability–precipitation combinations that occur in England; the remaining combination, high permeability with high precipitation, is essentially absent (0.3% of classified works) because permeable superficial and bedrock geology is concentrated in the drier south and east.

The fitted soil moisture, borehole groundwater level, dynamic composition and predicted versus simulated flow at the WWTP are shown in Fig. 6.

Westwood Brook WWTP is a high precipitation site upon mudstone, sandstone and siltstone bedrock with no superficial deposits, classified as a low permeability site. It has a balanced composition across sources, its model performance is in the top 1.5% of stations. Its municipal contribution varies between 20 and 100%, with some periods exhibiting an equal split between municipal, rainfall, phreatic, and vadose zone infiltration; the interpretation of any individual spill or sample at this site therefore depends strongly on its timing. It exemplifies a WWTP where composition analysis is of critical importance. Despite sewer overflows in this catchment spilling 88 times in 2025 for a total of 650 h³³, infiltration is not reported as a driver of overflow causes, illustrating the gap that composition analysis can fill.

Oakley WWTP is a high precipitation site upon clay, silt, sand and gravel superficial deposits above a chalk bedrock, classified as a low permeability site. Oakley demonstrates ephemeral phreatic zone infiltration particularly well, with groundwater overtopping into the sewer network in 2014. Any other variability is generally precipitation dominated and highlights the difficulty in achieving the highest, >0.75KGE, performance values in the wetter catchments, Fig. 1, which we attribute to the dominant influence of heterogeneous sewer infrastructure during rainfall events, and to storage within the works shifting event volumes between days.

Chesham WWTP is a low precipitation site upon chalk bedrock with no superficial deposits, classified as a high permeability site. It is also a site of special attention to the private water company that operates the plant owing to its widespread phreatic zone infiltration confirmed by field samples. The

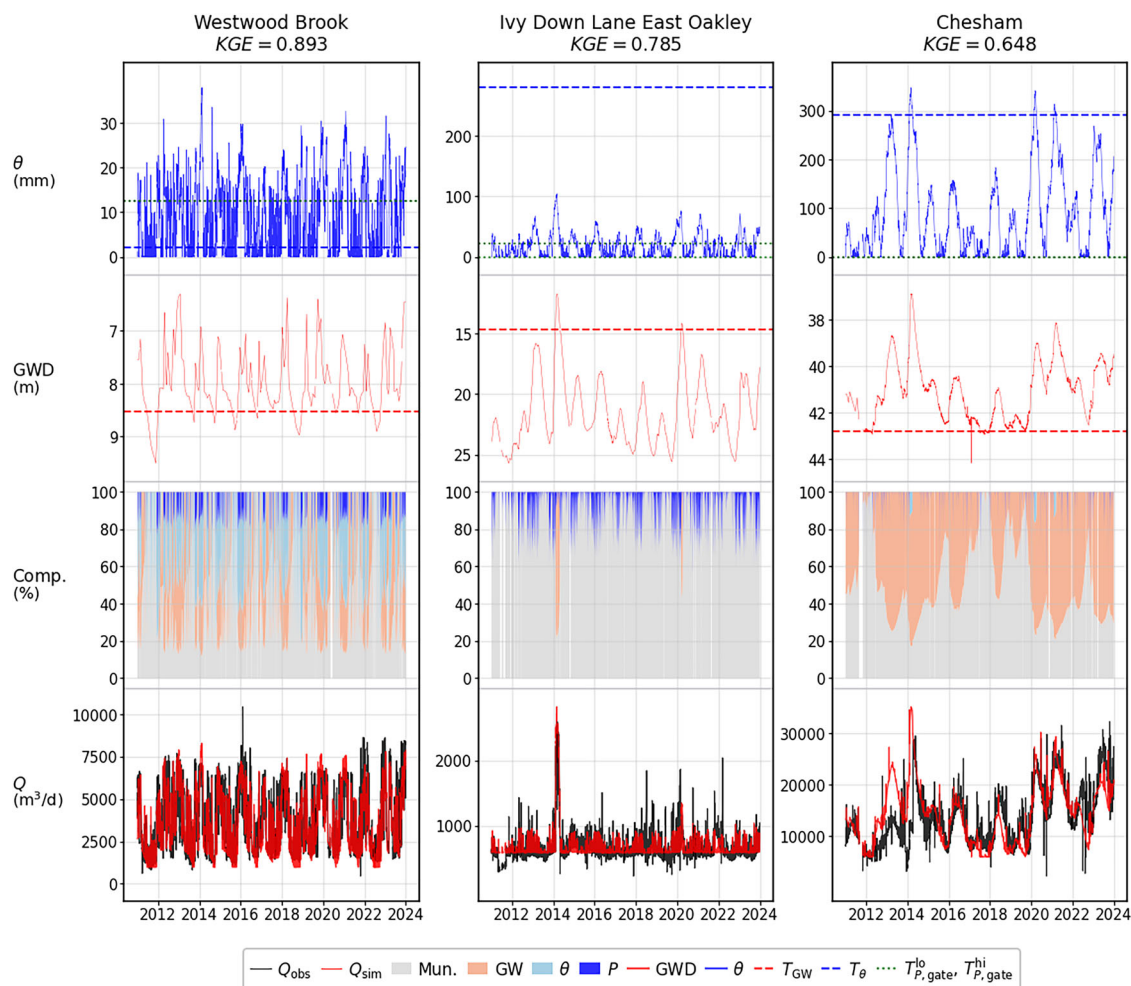


Fig. 6 | Selected site time series and inferred wastewater composition. Top row shows fitted soil moisture (θ), second row shows borehole observations formulated as groundwater depth below surface datum (GWD), third row shows dynamic composition of influent, bottom row shows predicted and observed WWTP flows.

WWTP flow variability tracks groundwater level in contrast to soil moisture, which instead has peaks occurring 3–6 months earlier than groundwater level peaks. The significant depths recorded at the assigned borehole (38–42 m) are deeper than any pipe would be set, indicating that the infiltration is likely occurring in a localised area, presumably adjacent to the rivers in the catchment; a single assigned borehole cannot be assumed representative of groundwater conditions across the catchment (see ‘Discussion’).

Discussion

Our central result is that source-resolved wastewater composition is recoverable at national scale from open data alone: WWCompose reproduces observed flows out of sample across a never before demonstrated number of WWTPs (Fig. 1), without explicit sewer-network topology. The composition maps, Fig. 3, reveal widespread non-municipal flow arriving at WWTPs. When disaggregated over time, for example, as shown in Westwood Brook, Fig. 6, the importance of wastewater decomposition becomes even more apparent, with non-municipal sources dominating during wetter periods. The spatial parameter patterns in Fig. 2 and detailed parametric assessment in Fig. 4 further indicate that the fitted quantities organise around hydrogeology, permeability, urbanisation, and soil in broadly intuitive relationships. Although WWTP archetypes were not apparent, the clearest catchment characteristics to delineate behavioural differences were aridity and permeability, supporting results from³⁴. Because no national open data on sewer asset age or condition exist, the fitted quantities are effective parameters that absorb local asset condition: at a given geology and

climate, higher groundwater or soil moisture composition is interpretable as a leakier network, so Fig. 2 can also be read as a screening indicator of relative network condition—an interpretation directly testable where utilities hold asset registers. The catchment controls (aquifer presence, permeability, aridity and urbanisation) covary with one another and with the geography of data availability such as usable boreholes—so disentangling their independent effects, together with spatial statistics for individual parameter fields (e.g. Moran’s I) and additional explanatory data such as sewer asset age and condition where these become available, is left to future work.

We identify a variety of use cases for WWCompose, discussed below.

Understanding the impact of combined sewer overflow spills is a key focus area of successive UK governments at present, who have committed £60 billion to their remediation³⁰. Existing literature has already indicated that 80% of total monitored spill duration is likely to be driven by soil moisture or groundwater sources⁴. Wastewater decomposition facilitates prioritising spills based on their proportion that is municipal, and targeting investment to the sites with most concentrated sewage spills. Furthermore, if a site has been selected for remediation, decomposition can direct appropriate solutions. For example, Nature Based Solutions (NBS) are a commonly proposed sewer spill intervention³⁵, and they may reduce spills that occur during high precipitation composition periods, however, they may worsen spills that occur during high soil moisture composition periods and will either not impact or potentially worsen, depending on their spatial extent, spills that occur during high groundwater composition periods.

A variety of research fields benefit from more granular detail in the influent to a WWTP. Provision of municipal composition is directly usable

for wastewater epidemiologists⁹, and, if paired with sample timings, may reduce the large uncertainty in interpretation attributable to the unknown human-waste fraction³⁶. In contrast, pairing composition with influent chemical samples may facilitate better quantifying diffuse urban pollution, which also has a significant uncertainty owing to the variability in composition of a given sample³⁷, and has already been demonstrated for extreme precipitation events in urban drainage systems³⁸. More widely, dynamic simulation of wastewater influent is required for analysis and design of WWTP technology³⁹. Phenomenological models, such as ref. 40 have demonstrated the value of splitting influent into different processes to generate such simulations. WWCompose profiles could be further developed to link to pollutant sources to predict water quality profiles that achieve similar dynamic water quality which could be used for WWTP performance analysis. For site-specific applications, the robustness of a decomposition could be further examined by sampling parameter sets whose objective lies within a practically indistinguishable band of the optimum and inspecting the spread of the resulting compositions. Water quality observations could also constrain the decomposition itself: conservative tracers such as electrical conductivity or chloride, interpreted through end-member mixing, distinguish municipal from groundwater and rainfall sources and could be added to the calibration objective to reduce equifinality where flow alone cannot. In England, routine regulatory samples are too sparse – typically monthly – to constrain a daily model, but targeted sampling at works experiencing equifinality (e.g., Chesham in Supplementary Fig. 1) may be effective.

Wastewater networks are the key driver of urban hydrology and are integral even at catchment scale hydrology⁴¹, interacting with fluvial and pluvial flooding. Identifying composition enables better quantification of wastewater processes in both hydrological and hydrodynamic pipe models. The aggregated conceptual representation may be broadly compatible with lumped hydrology models and could potentially be integrated into large scale hydrology initiatives such as CAMELS-GB⁴². For more detailed urban drainage models, results are partially transferable. Soil moisture representation in the Stormwater Management Model (SWMM)⁴³, a popular and open-source model for urban drainage modelling, includes a Green-Ampt infiltration model and power-law relationship for soil-to-pipe inflows as in WWCompose. Although this behaviour is defined in SWMM as a groundwater representation, based on the analysis presented in this work, it would more accurately be described to capture soil moisture-driven infiltration. The prevalence of groundwater contributions in some locations highlights the importance of enabling large scale aquifer groundwater table overtopping in SWMM-like models, which currently requires external packages to implement⁴⁴. As with sewer spills, the integrated picture provided by WWCompose may be used to guide flood interventions, helping to understand the interplay between precipitation-driven, soil moisture-driven, or groundwater-driven flooding.

WWCompose recovers composition from flow alone, and its main limitations follow from that design. Calibration validates the reproduction of total flow rather than of the individual components, so different parameter arrangements could in principle produce similar flows (equifinality). A small regularisation term resolves otherwise indistinguishable decompositions towards the more direct entry pathway (municipal over groundwater, rainfall over soil moisture; see Methods). Parameter equifinality arising due to differences in meteorological calibration period, measured by composition spread across repetitions, is generally small (see Results), however, in some cases, for example, where groundwater is dominant, it is difficult to distinguish from municipal, as illustrated in Chesham, Supplemental 1.

The municipal term is deliberately constant and combines both residential and non-residential wastewater. We avoided variation in residential water use to avoid conflation with the seasonal infiltration signals. Although population at wastewater catchment scale can be calculated⁴⁵, no national scale open data on non-residential wastewater could be sourced and so its inclusion would simply introduce an additional constant term into WWCompose and reduce parameter identifiability. Where composition

estimates are required for a specific catchment, local data on non-residential discharges should be sought; beyond this, population-presence modelling offers a demonstrated route to dynamic municipal estimates⁴⁶.

Performance reflects both data availability and applicability: held-out *KGE* increases with record length (Spearman $\rho = 0.33$), works size ($\rho = 0.34$) and catchment wetness ($\rho = 0.25$), factors that are only weakly inter-related (see Results). At small works in dry catchments the climate-driven signal that WWCompose recovers is weak relative to operational variability, which the model does not represent; longer records, spanning more hydrological variety, improve identifiability.

Effluent is used as a proxy for influent, requiring the assumption that the two are comparable at the daily timescale. Hydraulic residence time within larger works can exceed a day, and storm tanks and diurnal storage carry over volumes between days; these processes bias time-sensitive parameters—particularly the precipitation response—and contribute to the difficulty of fully capturing rainfall-driven dynamics even at otherwise well-performing wet sites such as Oakley. Influent gauging, where available, would remove this assumption.

Groundwater forcing relies on a single assigned borehole, which may not be representative of groundwater conditions across the catchment: at Chesham, borehole water-table depths of 38–42 m indicate that even field-confirmed chalk infiltration can be localised rather than network-wide. Further, 10% of works have no usable borehole, in which case the groundwater term is removed and any true phreatic contribution will be partly absorbed by the soil-moisture term, whose forcing co-varies with shallow groundwater.

Finally, parameters are fully site-specific. The framework transfers wherever comparable open data exist, but fitted values do not, and predicting composition at unmonitored works remains beyond the scope of this study.

In summary, national-scale wastewater composition can be inferred from open data, with 60% of English WWTPs exceeding a median held-out *KGE* of 0.6, in the largest such assessment demonstrated to date. The results show that non-municipal water is widespread even where municipal flow dominates long-term averages, and that permeability, groundwater, soil wetness, precipitation, and operational controls each leave identifiable signatures in effluent records. Aridity and permeability emerge as the clearest controls on composition, whereas distinct catchment archetypes do not. By turning a routinely monitored but rarely decomposed signal into a source-resolved, time-varying quantity, WWCompose provides a transferable basis for prioritising sewer overflow interventions, sharpening wastewater-based epidemiology, and representing wastewater within hydrological and urban drainage models. More broadly, it demonstrates that the dynamic provenance of wastewater is recoverable wherever comparable open data exist.

Methods

Our approach, called Wastewater Compose (WWCompose), extends existing empirical-conceptual studies such as the 'Triangle Method' and 'Moving Minimum' methods proposed by ref. 27 and 'Quasi-linear Model' approach proposed by ref. 25, which further provided dynamic estimates of dry-weather flow components. These studies recognise the difficulty in acquiring sufficient information to fully characterise a sewer network and instead create a conceptual model, splitting wastewater into constituent components, with few parameters that can be fitted to existing data. WWCompose, described in full in the Model structure section, assumes that the dynamic components of influent (rainfall, soil moisture, and groundwater) must exceed a specified threshold after which they contribute to influent with a linear and exponential component. We use climate data to force the model and, where possible, constrain these parameters with geospatial data, as described in the Data sources and alignment section. WWCompose is fit using best practice split-sample cross-validation and a sequential fitting process with final re-optimisation, described in the Calibration and validation section. Analyses of the fitted models are described in the Analysis of fitted models section.

Table 1 | Fitted parameters of WWCompose: physical meaning, units, optimisation bounds, and bound provenance

Symbol	Component	Physical meaning	Units	Bounds	Source of bounds
C_{Mun}	Municipal	Constant municipal (population and trade) flow	$m^3 d^{-1}$	$[0, 10^6]$	Exceeds largest observed flows
C_{GW}	Groundwater	Scaling of groundwater response	$m^3 d^{-1} \text{ per } m^{E_{GW}}$	$[10^{-6}, 10^4]$	Estimate
T_{GW}	Groundwater	Groundwater depth below which contributions begin	m	$[10^{-6}, 100]$	Spans observed borehole depth range
E_{GW}	Groundwater	Exponent of groundwater response	–	$[10^{-6}, 2]$	Sub- to super-linear response
C_θ	Soil moisture	Scaling of soil-moisture response	$m^3 d^{-1} \text{ per } mm^{E_\theta}$	$[10^{-6}, 10^4]$	Estimate
T_θ	Soil moisture	Soil-store depth threshold for contributions	mm	$[10^{-6}, 500]$	Spans feasible soil-store range
E_θ	Soil moisture	Exponent of soil-moisture response	–	$[10^{-6}, 2]$	Sub- to super-linear response
K_s	Soil moisture	Saturated hydraulic conductivity (Green–Ampt)	$mm d^{-1}$	site-specific	SoilGrids Q0.05–Q0.95
ψ	Soil moisture	Wetting-front suction head (Green–Ampt)	mm	site-specific	SoilGrids Q0.05–Q0.95
θ_s	Soil moisture	Saturated water content (Green–Ampt)	–	site-specific	SoilGrids Q0.05–Q0.95
θ_i	Soil moisture	Initial/field-capacity water content (Green–Ampt)	–	site-specific	SoilGrids Q0.05–Q0.95
C_P	Precipitation	Fraction of impervious-area rainfall excess reaching the works	–	$[0, 1]$	Runoff cannot exceed impervious-area rainfall volume
T_P	Precipitation	Daily rainfall threshold for runoff	$mm d^{-1}$	$[10^{-6}, 100]$	Spans daily rainfall range
T_P^{ub}	Precipitation	Cap on rainfall excess (drainage capacity)	$mm d^{-1}$	$[10^{-6}, 100]$	Spans daily rainfall range
E_P	Precipitation	Exponent of precipitation response	–	$[10^{-6}, 2]$	Sub- to super-linear response
$T_{P, \text{gate}}^{lo}$	Precipitation	Antecedent soil-moisture level at which the wetness gate opens	mm	$[0, 400]$	Spans feasible soil-store range
$T_{P, \text{gate}}^{hi}$	Precipitation	Antecedent soil-moisture level at which the gate saturates	mm	$[10^{-3}, 600]$	Spans feasible soil-store range

Bounds listed as site-specific are the SoilGrids Q0.05–Q0.95 uncertainty quantiles mapped to van Genuchten parameters via Rosetta, as described in Data sources and alignment.

Model structure

Simulated effluent flow at daily timestep t (days) is decomposed into municipal, groundwater, soil-wetness, and effective-precipitation terms, all in $m^3 d^{-1}$:

$$Q_{\text{sim}}(t) = C_{Mun} + Q_{GW}(t) + Q_\theta(t) + Q_P(t). \quad (1)$$

Besides C_{Mun} , which is assumed constant over the study period, each term is calculated by the distance a variable is above a threshold, however, the precise formulation differs for each term, as follows. Each dynamic component contributes through a shared threshold–scaling–exponent form (Eqs. (2), (6), (8)): this is the minimal form able to express threshold behaviour, magnitude, and non-linearity of response for each pathway, while remaining fittable from flow data alone. When bounds are specified below, such as $C_{Mun} \in [0, 1e6]$, these are the bounds within which the optimiser fits parameters, described in further detail in the Calibration and validation section; the fitted parameters, their units, bounds and bound provenance are summarised in Table 1. Although population and municipal water use vary over time, C_{Mun} is held constant for the reasons discussed in the Discussion.

Model structure: groundwater

Groundwater depth, $d(t)$ units (m), at the assigned borehole reduces groundwater contributions, and increases when the water table is shallow:

$$Q_{GW}(t) = C_{GW} \max(T_{GW} - d(t), 0)^{E_{GW}}, \quad (2)$$

with fitted coefficient $C_{GW} \in [0, 1e4]$, threshold $T_{GW} \in [0, 100]$, and exponent $E_{GW} \in [0, 2]$.

Model structure: soil moisture

A scalar soil-store depth $\theta(t)$ (mm) is forced by rainfall $P(t)$ and reduced by $ET_0(t)$, using a daily Green–Ampt-style recurrence⁴⁷, an established infiltration model with few parameters, all of which can be bounded by open soil data (Methods, Data sources and alignment). We use an explicit form to

improve computational efficiency during fitting:

$$\theta^*(t) = \max(\theta(t-1) - ET_0(t), 0), \quad (3)$$

$$f(t) = \begin{cases} K_s(\psi \Delta\theta + \varepsilon)/\varepsilon, & \theta^*(t) < \varepsilon, \\ K_s(1 + \psi \Delta\theta/\theta^*(t)), & \theta^*(t) \geq \varepsilon, \end{cases} \quad (4)$$

$$\theta(t) = \theta^*(t) + \min(P(t), f(t)). \quad (5)$$

Here $\Delta\theta = \theta_s - \theta_i$ with θ_s and θ_i dimensionless saturated water content and field capacity respectively; P , ET_0 , θ , and θ^* use depth units (mm). K_s is saturated hydraulic conductivity, ψ is wetting-front suction head, $f(t)$ is infiltration capacity, and ε is set to $1e-6$ to avoid division by 0. In the lower branch, ψ multiplies only the moisture deficit $\Delta\theta$, not cumulative storage θ^* . When $\theta^*(t) < \varepsilon$, the upper branch is a numerical guard; infiltration is then unconstrained by capacity ($f \gg P(t)$), so all rainfall infiltrates and $\theta(t) = \theta^*(t) + P(t)$. We fit K_s , ψ , θ_s , and θ_i within the soil probability bounds within WWTP catchment, calculated using SoilGrids (Methods, Data sources and alignment).

The soil-wetness contribution to flow is thus,

$$Q_\theta(t) = C_\theta \max(\theta(t) - T_\theta, 0)^{E_\theta}, \quad (6)$$

with fitted $C_\theta \in [0, 1e4]$, $T_\theta \in [0, 500]$, and $E_\theta \in [0, 2]$.

Model structure: precipitation

In addition to the generic value over threshold term, for example, Eq. (6), effective precipitation also uses capped rainfall excess intended to represent the capacity of the urban drainage system,

$$e_P(t) = \min\left(\max(P(t) - T_P, 0), \text{ub}T_P\right) \quad (7)$$

Furthermore, unlike the other components which cannot be meaningfully scaled, we multiply precipitation by the impervious area, A_{imp} (m^2), with rainfall converted from mm to m, calculated using^{32,48},

$$\tilde{Q}_p(t) = C_p A_{imp} e_p(t)^{E_p} \quad (8)$$

To capture the influence of semi-pervious surfaces reaching saturation and then contributing to precipitation-driven runoff, we include an interaction with antecedent soil moisture, $\theta_{lag}(t)$:

$$g_p(t) = \min \left(\max \left(\frac{\theta_{lag}(t) - \text{lo}T_{P, \text{gate}}}{\text{hi}T_{P, \text{gate}} - \text{lo}T_{P, \text{gate}}}, 0 \right), 1 \right) \quad (9)$$

The precipitation term is

$$Q_p(t) = g_p(t) \tilde{Q}_p(t). \quad (10)$$

Fitted parameters are $C_p \in [0, 1]$, $T_p \in [0, 100]$, $T_p^{\text{ub}} \in [0, 100]$, $E_p \in [0, 2]$, and the gate limits $T_{P, \text{gate}}^{\text{lo}} \in [0, 400]$, $T_{P, \text{gate}}^{\text{hi}} \in [0, 600]$.

Data sources and alignment

We use daily effluent flow records for English WWTPs from 2011 to 2023,³⁴ to fit WWCompose, requiring the assumption that, at a daily timescale, influent and effluent will be comparable. Diurnal variability and stormwater storage introduce errors here, discussed in the Discussion. A daily timestep is used because it is the common resolution of the national effluent and climate records, and because sub-daily influent variation is dominated by diurnal municipal patterns and the heterogeneity of sewer networks that WWCompose does not seek to resolve. The analysis only uses effluent observations certified as good quality in the raw data³⁴, via the per-observation quality flag assigned by the data provider.

HadUK gives catchment-mean daily rainfall⁴⁹, computed as the area-weighted average of the 1 km grid cells intersecting the catchment polygon (a catchment smaller than a single cell takes the value of the cell it intersects), which drives both the soil store and the excess-rainfall term (Methods, Model structure). ET_0 comes from HadUK via FAO-56 Penman-Monteith⁵⁰.

SoilGrids (sand, silt, clay, bulk density) at a fixed depth supplies the values at uncertainty quantiles (Q0.05, Q0.95)⁵¹. Rosetta maps each to van Genuchten parameters⁵²; these yield Green-Ampt K_s , ψ from K_s , θ_s , and θ_i (here at -33 kPa)⁴⁷. The quantiles define optimisation bounds for K_s , ψ , θ_s , and θ_i .

Borehole groundwater depths are retrieved from the English Environment Agency's hydrology data API, and assigned to each WWTP using the nearest borehole within 15 km. If the assigned series is absent or constant—281 works (10% of the cohort)—no groundwater term is present in Eq. (1); the consequence of this is discussed in the Discussion.

Permeability class K , used to interpret results, is assigned from superficial deposits where present and bedrock where absent, using expert knowledge informed by lithological descriptions of the British Geological Survey's 1: 50,000 scale superficial and bedrock geology maps.

Calibration and validation

Calibration fits daily simulated totals to observed effluent over the aligned window at each site within the study period 2011–2023 (Methods, Data sources and alignment). Parameters are estimated by maximising Kling-Gupta efficiency (KGE) on raw daily flows:

$$\text{KGE} = 1 - \sqrt{(r - 1)^2 + (\alpha - 1)^2 + (\beta - 1)^2}, \quad (11)$$

where r is correlation, α is the variability ratio, and β is the mean ratio between simulated and observed flows. KGE decomposes model error into correlation (timing), variability ratio (spread) and bias ratio (volume), and

maximising it constrains all three simultaneously⁵³. This matters for decomposition of WWTP influent in particular: the municipal component sets the volume of dry-weather flow, infiltration processes inflate the variance, and rainfall responses determine event timing – a credible decomposition must reproduce volume, spread and timing together. By contrast, R^2 measures the strength of linear association alone and is insensitive to systematic bias in volumes, while the Nash–Sutcliffe efficiency is dominated by high flows⁵³. In a catchment-scale study fitted to effluent records, such as WWCompose, high flows carry the largest errors: they are modified by storm storage and spills before reaching the works, and they are the most sensitive to sewer-network heterogeneity that the method does not represent. $\text{KGE} = 1$ indicates a perfect fit, and KGE values are not directly comparable to NSE or R^2 scales⁵⁴.

Each aligned record is partitioned into ten equal-length contiguous blocks; because record start and end dates differ between sites, block boundaries are site-specific. Contiguous blocks are used rather than random splits because they preserve the autocorrelation of daily flows, providing a fairer test of out-of-sample skill. For each repetition, one block is withheld for validation and parameters are trained on the remaining blocks. The held-out block moves across repetitions so that out-of-sample skill is summarised over different hydro-meteorological periods. Per-station skill is summarised by the median across blocks, and Fig. 1 shows the 25th–75th percentile band across replicates, because anomalous held-out periods can severely degrade an individual replicate.

WWCompose contains 17 parameters and the objective is non-convex, so we use sequential optimisation with joint refinement. Fitting proceeds in stages: the municipal constant and groundwater parameters are fitted first, then precipitation parameters, then soil-moisture parameters (including the Green–Ampt soil store). At each stage, parameters fitted earlier are frozen, the municipal constant is re-estimated, and components not yet introduced contribute zero flow. A final joint refinement then optimises all 17 parameters simultaneously. The objective at every stage is

$$L = -\text{KGE} + 10^{-3} \frac{\overline{Q_{\text{GW}}}}{\overline{Q_{\text{sim}}}} + 10^{-3} \frac{\overline{Q_{\theta}}}{\overline{Q_{\text{sim}}}}, \quad (12)$$

where the overbar denotes the mean over the training period and the two penalty terms favour municipal over groundwater and precipitation over soil moisture respectively; WWCompose prefers these because conventional knowledge suggests that they are the more direct routes for influent to enter the sewer network. The weights are small enough that the penalty is negligible relative to meaningful KGE differences and acts only to break ties between otherwise indistinguishable decompositions.

Every fitting stage uses differential evolution, a population-based global search, followed by local polishing of the best candidate (bound-constrained quasi-Newton refinement of the same objective); the same optimiser configuration is used at every stage, both per-component and joint. We use the implementation from Python SciPy⁵⁵. The results shown in this paper use a population size of 250 with maximum iterations of 3000, which testing revealed to provide converged objectives. These settings for WWCompose can be run for an individual WWTP on a MacBook Air M2 2022 in under 30 min, while the full set of experiments was performed using High-Performance Computing, see Acknowledgments.

Analysis of fitted models

A variety of data analyses were carried out to both verify the WWCompose data and understand the implications of its fitted parameters and results, which are explained in this section.

Performance is summarised per WWTP by the median KGE over the ten held-out cross-validation blocks; the median is used because anomalous held-out periods can severely degrade an individual replicate (Methods, Calibration and validation). Associations between performance and catchment characteristics (record length, works size, mean rainfall) are assessed by Spearman rank correlation.

Compositions are computed daily as each component's fraction of simulated flow. We present compositions both in their raw forms for exemplar catchments, and at national scale averaged over the record. For the national scale averaging, we weight all days equally rather than by total flow volume, adjudging this to be a more operationally useful value that is not dominated by high flow days. The sensitivity of the decomposition to the fitting period is summarised by the inter-quartile range of these average compositions across the ten cross-validation fits.

Stations are stratified for comparison. The stratum are the driest (and wettest) 5% stations whose average catchment daily rainfall lies below the 5th (and above the 95th) percentile across all stations; each group is compared against all remaining stations. The 'high permeability' classifications (see Data sources and alignment) are compared against low and medium combined. Stratum, S , compare the distribution of each variable x (a fitted parameter, average composition fraction, or held-out KGE) with a two-sample Kolmogorov–Smirnov statistic $D = \sup |F_S(x) - F_{\bar{S}}(x)|$ between the stratum and all remaining stations, and are reported signed by the direction of the mean difference: signed $D = \text{sign}(\text{mean}_S - \text{mean}_{\bar{S}}) \cdot D$.

Common structure across parameters is assessed by principal component analysis of the station-median values of all 17 fitted parameters (including those omitted from Fig. 2), standardised to zero mean and unit variance per parameter; cluster structure is assessed by k-means clustering of the same standardised values. Both use scikit-learn⁵⁶.

Data availability

The data generated by this study are available at <https://doi.org/10.5281/zenodo.21652986>. The study additionally uses openly available third-party data: daily effluent flow records for English WWTPs³⁴, HadUK-Grid climate data⁴⁹, SoilGrids soil properties⁵¹, borehole groundwater levels from the English Environment Agency hydrology data API, and British Geological Survey geology mapping³¹.

Code availability

The code that reproduces the full workflow (data cleaning, processing, simulation, analysis and visualisation) is available at <https://doi.org/10.5281/zenodo.21677607>.

Received: 15 June 2026; Accepted: 21 August 2026;

Published online: 04 September 2026

References

1. Staufer, P., Scheidegger, A. & Rieckermann, J. Assessing the performance of sewer rehabilitation on the reduction of infiltration and inflow. *Water Res.* **46**, 5185–5196 <https://doi.org/10.1016/j.watres.2012.07.001> (2012).
2. Becker, B., Reichel, F., Bachmann, D. & Schinke, R. High groundwater levels: processes, consequences, and management. *WIREs Water* **9** <https://doi.org/10.1002/wat2.1605> (2022).
3. Thapa, J. B., Jung, J. K. & Yovichin, R. D. A qualitative approach to determine the areas of highest inflow and infiltration in underground infrastructure for urban area. *Adv. Civil Eng.* <https://doi.org/10.1155/2019/2620459> (2019).
4. Lipp, A. G. & Dobson, B. Infiltration drives a heavy-tailed distribution of combined sewer overflow spill durations. *Environ. Res. Commun.* **8**, 041017 (2026).
5. Thorndahl, S., Schaarup-Jensen, K. & Rasmussen, M. R. On hydraulic and pollution effects of converting combined sewer catchments to separate sewer catchments. *Urban Water J.* **12**, 120–130 <https://doi.org/10.1080/1573062x.2013.831915> (2013).
6. Giakoumis, T. & Voulvoulis, N. The pollution load of combined sewer overflows and risks to England's waterbodies: relating event duration monitoring data to discharge consents from wastewater treatment works. *Environ. Sci. Water Res. Technol.* **12**, 1393–1412 (2026).
7. Carneiro, R. B. et al. A critical review of wastewater-based epidemiology as a tool to evaluate the unintentional human exposure to potentially harmful chemicals. *Anal. Bioanal. Chem.* **417**, 495–511 (2025).
8. Dirckx, G., Fenu, A., Wambecq, T., Kroll, S. & Weemaes, M. Dilution of sewage: Is it, after all, really worth the bother? *J. Hydrol.* **571**, 437–447 <https://doi.org/10.1016/j.jhydrol.2019.01.065> (2019).
9. Guerrero-Latorre, L. et al. The Catalan surveillance network of sars-cov-2 in sewage: design, implementation, and performance. *Sci. Rep.* **12**, 16704 (2022).
10. Pangle, L. A., Diem, J. E., Milligan, R., Adams, E. & Murray, A. Contextualizing inflow and infiltration within the streamflow regime of urban watersheds. *Water Resour. Res.* **58** <https://doi.org/10.1029/2021wr030406> (2022).
11. Karpf, C. & Krebs, P. Modelling of groundwater infiltration into sewer systems. *Urban Water J.* **10**, 221–229 <https://doi.org/10.1080/1573062x.2012.724077> (2013).
12. Sangsefidi, Y., Barnes, A., Merrifield, M. & Davani, H. Data-driven analysis and integrated modeling of climate change impacts on coastal groundwater and sanitary sewer infrastructure. *Sustain. Cities Soc.* **99**, 104914 <https://doi.org/10.1016/j.scs.2023.104914> (2023).
13. Ye, L., Qian, Y., Zhu, D. Z. & Huang, B. Inflow and infiltration assessment of a prototype sanitary sewer network in a coastal city in China. *Water Sci. Technol.* **88**, 2940–2954 <https://doi.org/10.2166/wst.2023.386> (2023).
14. Zeydilinejad, N., Javadi, A. A. & Webber, J. L. Global perspectives on groundwater infiltration to sewer networks: a threat to urban sustainability. *Water Res.* **262**, 122098 <https://doi.org/10.1016/j.watres.2024.122098> (2024).
15. Su, X., Liu, T., Beheshti, M. & Prigiobbe, V. Relationship between infiltration, sewer rehabilitation, and groundwater flooding in coastal urban areas. *Environ. Sci. Pollut. Res.* **27**, 14288–14298 <https://doi.org/10.1007/s11356-019-06513-z> (2019).
16. Zeydilinejad, N., Javadi, A. A., Jacob, M. & Webber, J. L. Evaluating groundwater infiltration risk in sewer networks under climate change. *Stoch. Environ. Res. Risk Assess.* **40**, 41 (2026).
17. Su, X., Belvedere, P., Tosco, T. & Prigiobbe, V. Studying the effect of sea level rise on nuisance flooding due to groundwater in a coastal urban area with aging infrastructure. *Urban Clim.* **43**, 101164 <https://doi.org/10.1016/j.uclim.2022.101164> (2022).
18. Wittenberg, H. & Aksoy, H. Groundwater intrusion into leaky sewer systems. *Water Sci. Technol.* **62**, 92–98 <https://doi.org/10.2166/wst.2010.287> (2010).
19. Martínez, C., Vojinovic, Z., Price, R. & Sanchez, A. Modelling infiltration process, overland flow and sewer system interactions for urban flood mitigation. *Water* **13**, 2028 <https://doi.org/10.3390/w13152028> (2021).
20. Zeydilinejad, N. An overview of the methods for evaluating the resilience of groundwater systems. *MethodsX* **10**, 102134 <https://doi.org/10.1016/j.mex.2023.102134> (2023).
21. Zhang, M. et al. Quantifying rainfall-derived inflow and infiltration in sanitary sewer systems based on conductivity monitoring. *J. Hydrol.* **558**, 174–183, <https://doi.org/10.1016/j.jhydrol.2018.01.002> (2018).
22. Dunning, C., Feinstein, D., Hunt, R. & Krohelski, J. *Simulation of Ground-water Flow, Surface-water Flow, and a Deep Sewer Tunnel System in the Menomonee Valley, Milwaukee, Wisconsin* <https://doi.org/10.3133/sir20045031> (2004).
23. Zeydilinejad, N. Artificial neural networks vis-à-vis MODFLOW in the simulation of groundwater: a review. *Model. Earth Syst. Environ.* **8**, 2911–2932 <https://doi.org/10.1007/s40808-022-01365-y> (2022).
24. Aßmann, E. et al. Augmentation of wastewater-based epidemiology with machine learning to support global health surveillance. *Nat. Water* **3**, 753–763 (2025).
25. Karpf, C. & Krebs, P. Quantification of groundwater infiltration and surface water inflows in urban sewer networks based on a multiple model approach. *Water Res.* **45**, 3129–3136 <https://doi.org/10.1016/j.watres.2011.03.022> (2011).

26. Dirckx, G., Bixio, D., Thoeye, C., De Gueldre, G. & Van De Steene, B. Dilution of sewage in Flanders mapped with mathematical and tracer methods. *Urban Water J.* **6**, 81–92 <https://doi.org/10.1080/15730620802541615> (2009).
27. Weiß, G., Brombach, H. & Haller, B. Infiltration and inflow in combined sewer systems: long-term analysis. *Water Sci. Technol.* **45**, 11–19 <https://doi.org/10.2166/wst.2002.0112> (2002).
28. Jenssen Sola, K., Bjerkholt, J. T., Lindholm, O. G. & Ratnaweera, H. Infiltration and inflow (i/i) to wastewater systems in Norway, Sweden, Denmark, and Finland. *Water* **10**, 1696 (2018).
29. Tuo, X. et al. Rapid on-site screening of excessive infiltration and inflow into sewer systems. *J. Hydrol. X* **31**, 100219 (2026).
30. Department for Environment, Food and Rural Affairs. Storm overflows discharge reduction plan. Tech. Rep., HM Government, London <https://www.gov.uk/government/publications/storm-overflows-discharge-reduction-plan> Expanded plan, updated 25 September 2023 (2023).
31. British Geological Survey. Hydrogeology 625k: digital hydrogeological map of the UK. <https://www.bgs.ac.uk/datasets/hydrogeology-625k/> Accessed July 2026.
32. Dobson, B., Jovanovic, T., Alonso-Álvarez, D. & Chegini, T. Swmm anywhere: a workflow for generation and sensitivity analysis of synthetic urban drainage models, anywhere. *Environ. Model. Softw.* **186**, 106358 (2025).
33. Lipp, A. G., Dawe, J. & Balaji, S. <https://www.sewagemap.co.uk> and poopy: open-source tools for understanding and communicating the environmental impacts of combined sewer overflows in real-time. *EarthArXiv* (2025).
34. Maes-Prior, R., Dobson, B. & Mijic, A. Drivers of wastewater dynamics: a statistical analysis of England's large wastewater treatment works. *Water Science and Technology* 295 (2026).
35. Moghanlo, S. J. & Raimondi, A. Impacts of blue-green infrastructures on combined sewer overflows. *Nat. Based Solut.* **7**, 100208 (2025).
36. van der Drift, A.-M. R. et al. Wastewater surveillance studies on pathogens and their use in public health decision-making: a scoping review. *Sci. Total Environ.* **993**, 179982 (2025).
37. Chen, Y., Yang, Z., Dong, J., Hong, N. & Tan, Q. Understanding phosphorus fractions and influential factors on urban road deposited sediments. *Sci. Total Environ.* **921**, 170624 (2024).
38. Hu, J., Ouyang, W., Hou, C., Ji, K. & Yang, Z. Effects of urban stormwater pollution on watershed diffuse loads under extreme precipitation conditions. *J. Hydrol.* **654**, 132802 (2025).
39. Martin, C. & Vanrolleghem, P. A. Analysing, completing, and generating influent data for WWTP modelling: a critical review. *Environ. Model. Softw.* **60**, 188–201 (2014).
40. Gernaey, K. V., Flores-Alsina, X., Rosen, C., Benedetti, L. & Jeppsson, U. Dynamic influent pollutant disturbance scenario generation using a phenomenological modelling approach. *Environ. Model. Softw.* **26**, 1255–1267 (2011).
41. Coxon, G. et al. Wastewater discharges and urban land cover dominate urban hydrology signals across England and Wales. *Environ. Res. Lett.* **19**, 084016 (2024).
42. Coxon, G. et al. CAMELS-GB: hydrometeorological time series and landscape attributes for 671 catchments in Great Britain. *Earth Syst. Sci. Data* **12**, 2459–2483, <https://doi.org/10.5194/essd-12-2459-2020> (2020).
43. Gironás, J., Roesner, L. A., Rossman, L. A. & Davis, J. A new applications manual for the storm water management model (SWMM). *Environ. Model. Softw.* **25**, 813–814, <https://doi.org/10.1016/j.envsoft.2009.11.009> (2010).
44. Zheng, Y., Chen, S., Qin, H. & Jiao, J. J. Modeling the spatial and seasonal variations of groundwater head in an urbanized area under low impact development. *Water* **10**, 803 (2018).
45. Hoffmann, T., Bunney, S., Kasprzyk-Hordern, B. & Singer, A. C. Wastewater catchment areas in Great Britain. *ESS Open Archive Preprint* at <https://doi.org/10.1002/essoar.10510612.2> (2025).
46. Dobson, B. et al. Integrated modelling to support analysis of COVID-19 impacts on London's water system and in-river water quality. *Front. Water* **3**, 641462 (2021).
47. Green, W. H. & Ampt, G. A. Studies on soil physics: Part I. The flow of air and water through soils. *J. Agric. Sci.* **4**, 1–24 <https://doi.org/10.1017/S0021859600001441> (1911).
48. Dobson, B., Alonso-Álvarez, D. & Chegini, T. Swmm anywhere: synthesise urban drainage network models anywhere in the world. *J. Open Source Softw.* **10**, 7729 (2025).
49. Hollis, D., McCarthy, M., Kendon, M., Legg, T. & Simpson, I. HadUK-Grid—a new UK dataset of gridded climate observations. *Geosci. Data J.* **6**, 151–159, <https://doi.org/10.1002/gdj3.78> (2019).
50. Allen, R. G., Pereira, L. S., Raes, D. & Smith, M. *Crop Evapotranspiration—Guidelines for Computing Crop Water Requirements—FAO Irrigation and Drainage Paper 56* FAO Irrigation and Drainage Paper No. 56 (Food and Agriculture Organization of the United Nations, 1998).
51. Poggio, L. et al. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *SOIL* **7**, 217–240 <https://doi.org/10.5194/soil-7-217-2021> (2021).
52. Schaap, M. G., Leij, F. J. & van Genuchten, M. T. ROSETTA: a computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. *J. Hydrol.* **251**, 163–176 [https://doi.org/10.1016/S0022-1694\(01\)00466-8](https://doi.org/10.1016/S0022-1694(01)00466-8) (2001).
53. Gupta, H. V., Kling, H., Yilmaz, K. K. & Martinez, G. F. Decomposition of the mean squared error and nse performance criteria: Implications for improving hydrological modelling. *J. Hydrol.* **377**, 80–91 (2009).
54. Knoben, W. J. M., Freer, J. E. & Woods, R. A. Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrol. Earth Syst. Sci.* **23**, 4323–4331 (2019).
55. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
56. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

Acknowledgements

We acknowledge computational resources and support provided by the Imperial College Research Computing Service (<https://doi.org/10.14469/hpc/2232>).

Author contributions

B.D. and C.J. conceived the research questions, study design, and methods. B.D., C.J. and R.M-P. curated data, interpreted results, and contributed to writing the manuscript. B.D. run and analysed experiments and supervised the work.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41545-026-00630-2>.

Correspondence and requests for materials should be addressed to Barnaby Dobson.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.

© The Author(s) 2026