Title
Water balance creates a threshold in soil pH at the global scale

Permalink
https://escholarship.org/uc/item/30f631wk

Journal
Nature, 540(7634)

ISSN
0028-0836

Authors
Slessarev, EW
Lin, Y
Bingham, NL
et al.

Publication Date
2016-12-22

DOI
10.1038/nature20139

Peer reviewed
Water balance creates a threshold in soil pH at the global scale

E.W. Slessarev1, Y. Lin2, N. L. Bingham3, J. E. Johnson4, Y. Dai5, J. P. Schimel1 & O. A. Chadwick1

Soil pH regulates the capacity of soils to store and supply nutrients, and thus contributes substantially to controlling productivity in terrestrial ecosystems. However, soil pH is not an independent regulator of soil fertility—rather, it is ultimately controlled by environmental forcing. In particular, small changes in water balance cause a steep transition from alkaline to acid soils across natural climate gradients. Although the processes governing this threshold in soil pH are well understood, the threshold has not been quantified at the global scale, where the influence of climate may be confounded by the effects of topography and mineralogy. Here we evaluate the global relationship between water balance and soil pH by extracting a spatially random sample (n = 20,000) from an extensive compilation of 60,291 soil pH measurements. We show that there is an abrupt transition from alkaline to acid soil pH that occurs at the point where mean annual precipitation begins to exceed mean annual potential evapotranspiration. We evaluate deviations from this global pattern, showing that they may result from seasonality, climate history, erosion and mineralogy. These results demonstrate that climate creates a nonlinear pattern in soil solution chemistry at the global scale; they also reveal conditions under which soils maintain pH out of equilibrium with modern climate.

Climate controls many aspects of soil chemistry, affecting soil pH (ref. 4). Alkaline soils are known to be common in arid climates, while acid soils are known to be common in humid climates. Surprisingly, however, the global-scale mechanisms governing this pattern remain broadly defined, and untested by direct observation. What are the dominant chemical equilibria that constrain soil pH? What aspect of climate defines the transition between alkaline and acid soils, and is the transition linear? The answers to these questions are fundamental to understanding soil development and surface geochemistry at the global scale. Furthermore, achieving this understanding may prove essential for representing soils in models of the terrestrial biosphere, given that soil pH controls many aspects of soil fertility. Here we illustrate that simple geochemical and hydrological concepts can be used to build a mechanistic understanding of soil pH at the global scale.

Interpretations of acid–titration experiments indicate that the soil pH is typically most strongly buffered by equilibrium with two secondary minerals: calcite (CaCO$_3$) or gibbsite (Al(OH)$_3$)2. CaCO$_3$ precipitates from calcium ions (Ca$^{2+}$) and carbonate ions (CO$_3^{2-}$) derived from dissolved carbon dioxide (CO$_2$). Al(OH)$_3$ precipitates from aluminium ions (Al$^{3+}$) that are released from negatively charged exchange sites. Both CaCO$_3$ and Al(OH)$_3$ consume protons (H$^+$) when they dissolve and release H$^+$ when they precipitate, buffering soil pH (ref. 5). Under typical laboratory conditions, soils in equilibrium with CaCO$_3$ and atmospheric CO$_2$ have a pH of 8.2 (see Methods), while the pH of soils that contain exchangeable Al$^{3+}$ is on average 5.1 (see Methods). The presence of CaCO$_3$ and Al(OH)$_3$ is reflected in soil pH across a wide range of CaCO$_3$ and exchangeable Al$^{3+}$ concentrations (Extended Data Fig. 1).

Local studies of climate gradients have shown that the relative importance of these two buffers is determined by leaching, which removes Ca$^{2+}$ from the soil2,3,8,9. In climates where evaporative demand exceeds precipitation, leaching rates are low, and dissolved Ca$^{2+}$ accumulates as CaCO$_3$—buffering soil pH near 8.2 (ref. 4). Conversely, in climates where precipitation exceeds evaporative demand, water leaks through the soil, removing Ca$^{2+}$ and allowing accumulation of relatively immobile Al$^{3+}$—buffering soil pH near 5.1 (ref. 4). Because runoff and leaching rates increase abruptly as precipitation exceeds evaporative demand10, the transition between CaCO$_3$ and Al(OH)$_3$ buffered conditions is expected to occur over a small range of climatic forcing, creating a steep threshold in soil pH at the transition point between arid and humid climates.1,8

However, whereas leaching controls the loss rate of Ca$^{2+}$, topography and mineralogy control the supply rate of Ca$^{2+}$ to the soil solution via erosion and weathering11, and thus interact with climate to influence soil pH over long timescales. For instance, calcium-containing minerals may be terminally depleted in old, low-relief landscapes that have been leached in the past, limiting Ca$^{2+}$ supply to the soil solution and creating Al(OH)$_3$-buffered soils under arid conditions. Alternatively, soils with high residence times in steep landscapes or areas dominated by Ca-rich rock can be rapidly supplied with Ca$^{2+}$ from weathering, counteracting the accumulation of exchangeable Al$^{3+}$ (ref. 12). Variation in the Ca$^{2+}$ supply rate is challenging to constrain at global scales, and might obscure the fundamental relationship between climate and soil pH. Thus, it is unclear whether the threshold in soil pH predicted by theory prevails globally.

Nonetheless, we can search for the pH threshold at the global scale, given sufficiently extensive sampling. Statistically derived soil maps provide a tempting tool for validation13,14. However, these maps rely on spatial projections of soil taxonomy that are sometimes explicitly defined by climate15, and would provide circular evidence. Thus, to test our hypotheses, we used actual measurements sampled from public databases of soil profiles (Extended Data Table 1). We then focused on pH in the subsoil (assigned here as soil to a depth of 0.5 m), to avoid effects of land-use and vegetation that might obscure the underlying geochemical signal. To overcome spatial biases in the databases (for example, heavy sampling in the USA), we developed a simple re-sampling approach that selects soil profiles randomly with respect to geographic space (Extended Data Fig. 2). We then associated these pH measurements with 1° gridded estimates of mean annual precipitation (MAP)16 and a model of potential evapotranspiration (PET)17, which represents evaporative demand. This allowed us to separate water-limited climates where leaching rates are low (MAP minus PET < 0) from energy-limited climates where leaching rates are high (MAP minus PET > 0).

 Globally, the relationship between soil pH and MAP minus PET conforms closely to predictions. Soil pH at 0.5 m depth has two modes that approximate 8.2 and 5.1, the values associated with CaCO$_3$ and Al(OH)$_3$ buffers (Fig. 1). Where MAP minus PET approaches 0, there
soils (8.2) and Al(OH)$_3$-buffered soils (5.1). Fewer deviations towards alkaline pH exist in humid climates. However, some humid soils have pH values exceeding 6.5. These measurements are scattered across several regions, including (6) southern China, (7) northern and central Europe, (8) northeastern North America, and (9) and (10) the Pacific Rim (Fig. 2). In regions (6) to (8), carbonate rocks are a major component of bedrock lithology and (9) and (10), active volcanoes may produce easily weathered silicate minerals that could buffer pH outside the range of Al(OH)$_3$ equilibria. More generally, we observe that humid-climate soils are less acidic in high-relief landscapes (Extended Data Fig. 7), where high soil production rates may increase the availability of fresh Ca-containing minerals, increasing Ca$^{2+}$ supply to the soil solution and thus counteracting accumulation of exchangeable Al$^{3+}$.

Intriguingly, the bimodal shape of the soil pH distribution indicates that soils in the neutral pH range (pH 6–7) are uncommon relative to soils in the CaCO$_3$ and Al(OH)$_3$ buffered ranges. Soils in this pH region are thought to be buffered by mineral weathering reactions. In theory, the capacity of these reactions to neutralize H$^+$ is limited by the relatively slow kinetics of primary mineral dissolution, and so neutral-range soils may evolve towards CaCO$_3$ and Al(OH)$_3$ equilibria over time. Not coincidentally, neutral-range soils are intensively cultivated, because they cluster in sub-humid climates with sufficient rainfall.
for agriculture, but retain nutrients more effectively than acid soils. Thus, from an observational standpoint, the most naturally fertile soils are relatively uncommon—and hypothetically, their relatively low prevalence may result from intrinsic aspects of their chemistry.

By assuming a threshold between CaCO$_3$-buffered and Al(OH)$_3$-buffered domains where $MAP = PET$, we can explain 42% (interquartile range 42%–45%) of the global variation in soil pH. The strength of this pattern indicates that a small number of specific chemical and physical mechanisms govern soil pH at the global scale. Moreover, by using this pattern as a guide, we can identify soils that appear out of equilibrium with modern climate. The distribution of these soils suggests a range of new questions that apply to the timescales of soil development: are acid soils in arid, low-relief environments irreversibly leached? Can erosion maintain high pH at a steady state in humid climates? And are neutral-range soils less common because they are poorly buffered? The answers to these questions are relevant at the timescale of human societies. Rapid changes in water balance caused by climate or land-use change might leave an increasing number of soils out of equilibrium with climate, with unknown consequences for their capacity to support productivity in natural and managed ecosystems.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 25 April; accepted 22 September 2016.**

**Published online 21 November 2016.**


**Acknowledgements** We thank P. Vitousek, S. Fendorf, X. Feng and C. Koura for guidance and comments. Soil data were provided by multiple contributing organizations (Extended Data Table 1). Funding for E.W.S. was provided by a Graduate Research Fellowship from the United States National Science Foundation.

**Author Contributions** Research was conceived by E.W.S., O.A.C., Y.L., N.L.B., J.P.S. and J.E.J. Data aggregation and processing tasks were shared by E.W.S., Y.L. and J.E.J. Funding for E.W.S. was provided by a Graduate Research Fellowship from the United States National Science Foundation.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.W.S. (eric.slessarev@lifesci.ucsb.edu).

**Reviewer Information** Nature thanks R. Merckx, S. Porder and the other anonymous reviewer(s) for their contribution to the peer review of this work.
Evaporation models. We use potential evapotranspiration (PET) to represent the evaporative component of climate rather than actual evapotranspiration (AET), because PET is independent of precipitation, and thus carries more information about arid climates. Specifically, we reason that climates that are close to the MAP = PET transition are more likely to have been leached in the recent past than climates where PET greatly exceeds MAP, even if both climates have comparably small modern values of MAP minus AET. In this sense, modern PET may provide a better index of long-term leaching rates than modern AET.

We represented PET using two contrasting approaches. In the first approach, we represented PET using a modified Penman–Monteith–Leuning model, which estimates evaporation as a function of net radiation ($R_n$), air temperature, vapour-pressure deficit, and the aerodynamic and surface conductance of vegetation.17 This approach is biophysically detailed, but it requires many parameters. Thus, in the second approach, we represented PET using the comparatively simple Priestley–Taylor equation, which models evaporative demand as a function of net radiation, air temperature, and a scaling parameter, $\alpha$ (ref. 29). We also explored the relationship between soil pH and the difference of MAP and AET, which we hypothesized was influenced by windspeed and vegetation cover, and then identified all 1° cells that contained at least one 0.5° cell classified using the diagnostic data included in the LandFlux-EVAL. By definition, MAP minus AET cannot take negative values, but approaches zero where PET > MAP10. We report values of MAP minus AET without imposing this constraint, and so some modelled AET values exceed MAP, resulting in slightly negative values.

Modified Penman–Monteith–Leuning model. The Penman–Monteith–Leuning model partitions evaporation from the plant canopy ($E_C$) and soil ($E_S$). $E_C$ is estimated using the Penman–Monteith equation16, while evaporation from soil is assumed to equal the equilibrium rate, modified by a moisture constraint. Because we were interested in obtaining an estimate of PET, we did not include a soil moisture constraint on evaporation, and then assumed that PET was equal to the sum of canopy and soil evaporation. Evaporation from wet soil can be approximated by multiplying the equilibrium evaporation rate by the Priestley–Taylor coefficient, $\alpha$ (ref. 31). Thus, we substituted the Priestley–Taylor model for the equilibrium model to represent soil evaporation in the Penman–Monteith formulation. The combined evaporation from canopy and soil are given by the equation:

$$\lambda E_{\text{tot}} = (sA_c + \rho c_p D_G(G_s)/(s + \gamma (1 - G_s/G_c)) + \alpha (sA_s)/(s + \gamma)$$

(1)

where $A_c$ and $A_s$ are the available energy absorbed by canopy and soil ($R_n$ minus soil heat flux, in units of MJ m$^{-2}$ d$^{-1}$), $\lambda$ is the latent heat of vaporization of water (MJ kg$^{-1}$), $\rho$ is the density of air (kg m$^{-3}$), $c_p$ is the specific heat of air at constant pressure (MJ kg$^{-1}$ K$^{-1}$), $D_G$ is the vapour pressure deficit (kPa), $G_c$ is the aerodynamic conductance (m$^{-1}$ d$^{-1}$), and $G_s$ is the canopy conductance (m$^{-1}$ d$^{-1}$). Radiation is partitioned between canopy and soil by two equations:

$$A_c = A_{\text{tot}} (1 - e^{-k_L})$$

(2)

$$A_s = A_{\text{tot}} - A_c$$

(3)

where $A_{\text{tot}}$ is equal to $R_n$ (the soil heat flux is assumed to be negligible), $L$ is the leaf area index, and $k_L$ is an extinction coefficient. Canopy conductance ($G_c$) is constrained by maximum stomatal conductance ($g_{\text{max}}$), and modified by factors that represent dependence on light availability and vapour-pressure deficit:

$$G_c = (g_{\text{max}}/k_L)\ln[(Q_e + Q_a)/(Q_e - (k_QD/Q_a)(1/(1 + D_O/D_S))]$$

(4)

where $Q_e$ is photosynthetically active radiation at the top of the canopy (half of the incoming shortwave radiation), $Q_a$ is a half-saturation constant for $Q_e$, $D_O$ is a half-saturation constant for vapour-pressure deficit, and $k_Q$ is the extinction coefficient for short-wave radiation.

Most parameters were obtained from a regional implementation of the Penman–Monteith–Leuning model.22 The parameters $k_L$ and $k_Q$ were both set equal to 0.6 m$^{-1}$, while $Q_o$ and $D_S$ were set equal to 2.6 MJ m$^{-2}$ d$^{-1}$ and 0.8 kPa (ref. 32). The maximum stomatal conductance, $g_{\text{max}}$, was set equal to 0.006 m$^{-1}$ s$^{-1}$, which is a reasonable mean estimate for natural vegetation,15 and scaled to a daily time step. The aerodynamic conductance, $G_c$, is influenced by windspeed and vegetation height. Because reliable maps of both these parameters are unavailable at a global scale, we used biome-specific parameters, assigning forests and savannas a value of 0.033 m s$^{-1}$, shrublands a value of 0.0125 m s$^{-1}$, and grasslands, cropland, and barren areas 0.011 m s$^{-1}$. All other parameters were calculated or obtained from the Food and Agriculture Organization (FAO) guidelines.

Priestley–Taylor model. The Priestley–Taylor model for PET uses a single parameter, $\alpha$, to account for adiabatic component of latent heat transfer.24 While $\alpha$ may vary as a function of meteorological conditions, a standard $\alpha$ value of 1.26 has been applied successfully at large scales.53 Priestley–Taylor PET is given by the equation:

$$E_{\text{PET}} = a\lambda A/(s + \gamma)$$

(5)

where $A$ is total available energy (equal to $R_n$) and $\alpha = 1.26$. Other parameters are listed above.

Precipitation dataset. We estimated mean annual precipitation (MAP) using a 1° gridded map created from the Global Precipitation Climatology Center (GPM) data, which we used as the mean annual sum of monthly precipitation values for the years 1961–2001. We use this 40-year interval because it includes a high spatial coverage of rain-gauge stations.55 We corrected for systematic rain gauge measurement error using static monthly under-catch corrections provided by the Global Precipitation Climatology Center.

Driving data for PET. Both Penman–Monteith–Leuning and Priestley–Taylor models require monthly estimates of $R_n$ and air temperature, and the Penman–Monteith–Leuning model requires monthly estimates of vapour pressure, atmospheric pressure, surface short-wave radiation, leaf-area index, and land cover type (Extended Data Table 2). For both approaches, environmental variables obtained for multi-year time series were collapsed to monthly means of daily values before calculation of PET. PET was scaled from daily to monthly values, then summed to obtain annual PET. To test the sensitivity of our results to driving data, we used two radiation data sets: mean monthly values from the NASA/CERES energy-balanced and filled surface radiation budget, version 2.8, over the years 2001–2014,56 and mean monthly values from the NASA/GEWEX surface radiation budget version 3.0 over the years 1984–2007.57 We obtained mean monthly values of air temperature and vapour pressure from the CRU TS3.13 data set, a gridded climatology at 0.5° resolution interpolated from weather station measurements, which we averaged over the period 1961–2001, the period of maximum weather station coverage.58 Atmospheric pressure was obtained using mean elevations from theETOPO1 global digital elevation model59 in each 1° cell and correcting using the ideal gas law.60 Land cover classes were obtained from the NASA MODIS satellite product MOD1262 and monthly means of leaf area index for the period 2001–2012 were obtained from the MODIS-derived Global Land Surface Satellite leaf area index data set61,62, averaged over the period 2001–2012. All data at a higher resolution than 1° were aggregated to mean values at 1° resolution before calculation of PET.

Rainfall seasonality. We quantified rainfall seasonality by computing the coefficient of variation of under-catch corrected monthly rainfall values from the Global Precipitation Climatology Center data set.

Local relief. We estimated local topographic relief from the 1-arcminute resolution ETOPO1 digital elevation model.61 Local relief was calculated as the difference between maximum and minimum elevations within a 10-km radius of each 1° arcminute cell. Local relief at 1° resolution was then calculated as the median relief within each 1° cell.

Carbonate lithology. We represented the extent of carbonate lithology using the Global Lithologic Map (GLIM)63. We determined which 1° grid cells contained carbonate rocks by subsetting the 0.5° raster version of GLIM for carbonate lithology, and then identifying all 1° cells that contained at least one 0.5° cell classified as carbonate rock.

Soil profile data. We combined data from eight soil profile databases (Extended Data Table 1).45–48 Profiles were included if they were non-duplicated and included measurements of pH in soil-water suspension. We used pH in water rather than pH in CaCl$_2$ or KCl solutions because pH in water is reported at a much higher frequency than pH in salt solutions. Data at 0.5 m and 0.1 m depth were obtained by selecting the horizon of each profile intersected by the corresponding depth. We selected absolute depths at 0.5 m and 0.1 m rather than soil horizons because horizon nomenclature varied across data sets. Although the choice of depths is somewhat arbitrary, the depths were selected to span the depths at which biological cycling typically influences cation concentrations. Using the National Cooperative Soil Survey (NCSS) database as a reference, 0.5 m approximates the median value for the top of the B horizon (0.52 m) and 0.1 m approximates the median value for the midpoint of the A horizon (0.09 m). The total number of profiles included was 60,291 at 0.5 m depth and 67,900 at 0.1 m depth (Extended Data Table 1).

Dilution ratio correction. The soil-to-water ratio of the slurry used to measure soil pH varied across data sets. To account for the effects of the soil-to-water ratio, data reported for a 1:5 ratio were corrected to a 1:1 ratio using linear correction factors.65 We could not obtain correction factors for data measured at a ratio of 1:2.5, and so left these data uncorrected. Including uncorrected data is unlikely to drive large errors in the global pH distribution because changing the soil-to-water ratio from 1:1 to 1:5 shifts pH by about 0.5 units66, which is small relative to the
global range of soil pH values. Data measured in water without a reported ratio were assumed to be measured at ratios of 1:1 or 1:2.5.

### Statistical analyses

Soil profile data were spatially resampled. In this approach, individual soil profiles were selected based on proximity to randomly distributed sampling nodes (n = 20,000). Sampling nodes were drawn from grid-cell centres at 1° resolution, with sampling weights based on cell area and allowing replacement. Nodes that were more than 100 km from a soil profile were not sampled to minimize edge biases. Soil profiles were selected by identifying the closest grid cell to each atmospheric CO2 leveller, and then randomly selecting a profile from the total set of profiles in the cell. By design, this approach includes individual profiles multiple times in the resampled data set, with the consequence that geographically isolated profiles are included more frequently than profiles in densely sampled areas. This approach has no statistical derivation, but it produces a sampling distribution that appears less-biased than the underlying data (Extended Data Fig. 1).

### Water-balance model evaluation

To evaluate the relationship between MAP minus PET and soil pH, we compared observations to theoretical predictions based on calcite and gibbsite buffering systems. For all soils in grid cells where MAP minus PET < 0, the predicted pH was 8.2, and for all remaining profiles, the predicted pH was 5.1. Residuals from the model were then computed by subtracting predicted values from observed values. Because the data are bimodally distributed, residuals from this model have a heavy-tailed distribution, and measures of variation based on squared errors (for example, the coefficient of determination, R2) are inappropriate. Instead, we estimated variation in the data using a robust measure of dispersion, the median absolute difference from the median (MAD). We then gauged model fit by comparing the MAD of the residuals to the MAD of the data: the percentage variation explained was equal to 1 minus MADresidual/MADdata.

This metric is analogous to R2, but makes no assumption about the distribution of the data or residuals. We estimated the uncertainty in the percentage of variation explained by resampling the data with replacement 10,000 times and calculating the interquartile range of the resulting distribution of parameter estimates.

### Logistic regression of outliers

We defined ‘outliers’ as soils with pH < 6.5 in strongly arid climates (driest quartile of MAP minus PET) and soils with pH > 6.5 in strongly humid climates (wettest quartile of MAP minus PET). We deliberately reduced pH to this categorical expression to emphasize large-scale deviations between pH modes, rather than small-scale deviations around each mode. To quantify the prevalence of outliers as a function of rainfall seasonality, carbonate lithology, and topographic relief, we fitted logistic regressions. Likelihood ratio tests were used to compare regressions against the null hypothesis that the proportion of outliers is uniform with respect to each predictor. We ruled out possible collinearity between environmental predictors by checking individual correlations between predictors in both wet and dry climates. No two predictors had correlation coefficients above 0.25, and so we assume that the patterns presented are independent.

### Calcite and aluminum chemistry

We used the NCSS database to validate chemical calculations and determine the relationship between calcite (CaCO3), and exchangeable aluminum (Alx) in soils. The NCSS database for this purpose because it contains a large number of measurements of CaCO3 and Alx using consistent methods, and it reports the effective cation exchange capacity, which is required for modelling the pH of gibbsite buffered soils. We used a spatially resampled subset of 20,000 data points for plotting relationships with the annual water balance, following the resampling method above.

### Gibbsite buffer

The pH of a solution exposed to calcite (CaCO3) and open to the atmosphere can be solved using an equation derived from the chemical equilibrium for CaCO3 (ref. 56):

\[
0 = H^+ (K_h/K_sK_h + PCO_2) + H^+(HK_sK_h - PCO_2 - K_hK_h - K_hPCO_2)
\]

where \(H^+\) is the hydrogen ion activity in moles, \(K_h\) is the solubility constant for CaCO3 (in units of \(\text{mol}^{-1}\)), \(K_s\) is the dissociation constant for water (in \(\text{mol}^{-1}\)), \(K_h\) and \(K_s\) are the first and second dissociation constants of carbonic acid (in \(\text{mol}^{-1}\)), \(K_h\) is Henry’s constant (in \(\text{mol}^{-1}\)), and \(PCO_2\) is the partial pressure of CO2 (in atm). We solved this equation for \(H^+\) at 25°C and a \(PCO_2\) of 3.45 × 10^-4 atm using the package rootSolve in R and published parameters. The \(PCO_2\) value of 3.45 × 10^-4 atm reflects \(PCO_2\) imposed by laboratory measurement conditions at standard atmospheric pressure, based on the ambient CO2 mole fraction in 1986, the median measurement date of the data. Older measurements made at lower atmospheric CO2 levels may reflect a slightly higher calcite equilibrium pH (that is, the expected pH is 8.3 before 1977). Because this difference is small and the majority of measurements were taken after this date, we report model fits for a single \(PCO_2\) value.

Calcite concentrations are approximate, and reported as CaCO3 equivalents. The NCSS database reports CaCO3 equivalents measured using a pressure calicimeter following acid dissolution, meaning that a range of carbonate minerals are included in the estimate. Also, because values are reported at a precision of 1%, some soils with <1% CaCO3 are probably reported with zero values, even if their pH reflects buffering by CaCO3.

### Global range of soil pH values

Data measured in water without a reported ratio were assumed to be measured at ratios of 1:1 or 1:2.5.

### Statistical analyses

Soil profile data were spatially resampled. In this approach, individual soil profiles were selected based on proximity to randomly distributed sampling nodes (n = 20,000). Sampling nodes were drawn from grid-cell centres at 1° resolution, with sampling weights based on cell area and allowing replacement. Nodes that were more than 100 km from a soil profile were not sampled to minimize edge biases. Soil profiles were selected by identifying the closest grid cell to each atmospheric CO2 leveller, and then randomly selecting a profile from the total set of profiles in the cell. By design, this approach includes individual profiles multiple times in the resampled data set, with the consequence that geographically isolated profiles are included more frequently than profiles in densely sampled areas. This approach has no statistical derivation, but it produces a sampling distribution that appears less-biased than the underlying data (Extended Data Fig. 1).

### Water-balance model evaluation

To evaluate the relationship between MAP minus PET and soil pH, we compared observations to theoretical predictions based on calcite and gibbsite buffering systems. For all soils in grid cells where MAP minus PET < 0, the predicted pH was 8.2, and for all remaining profiles, the predicted pH was 5.1. Residuals from the model were then computed by subtracting predicted values from observed values. Because the data are bimodally distributed, residuals from this model have a heavy-tailed distribution, and measures of variation based on squared errors (for example, the coefficient of determination, R2) are inappropriate. Instead, we estimated variation in the data using a robust measure of dispersion, the median absolute difference from the median (MAD). We then gauged model fit by comparing the MAD of the residuals to the MAD of the data: the percentage variation explained was equal to 1 minus MADresidual/MADdata.

This metric is analogous to R2, but makes no assumption about the distribution of the data or residuals. We estimated the uncertainty in the percentage of variation explained by resampling the data with replacement 10,000 times and calculating the interquartile range of the resulting distribution of parameter estimates.

### Logistic regression of outliers

We defined ‘outliers’ as soils with pH < 6.5 in strongly arid climates (driest quartile of MAP minus PET) and soils with pH > 6.5 in strongly humid climates (wettest quartile of MAP minus PET). We deliberately reduced pH to this categorical expression to emphasize large-scale deviations between pH modes, rather than small-scale deviations around each mode. To quantify the prevalence of outliers as a function of rainfall seasonality, carbonate lithology, and topographic relief, we fitted logistic regressions. Likelihood ratio tests were used to compare regressions against the null hypothesis that the proportion of outliers is uniform with respect to each predictor. We ruled out possible collinearity between environmental predictors by checking individual correlations between predictors in both wet and dry climates. No two predictors had correlation coefficients above 0.25, and so we assume that the patterns presented are independent.

### Calcite and aluminum chemistry

We used the NCSS database to validate chemical calculations and determine the relationship between calcite (CaCO3), and exchangeable aluminum (Alx) in soils. The NCSS database for this purpose because it contains a large number of measurements of CaCO3 and Alx using consistent methods, and it reports the effective cation exchange capacity, which is required for modelling the pH of gibbsite buffered soils. We used a spatially resampled subset of 20,000 data points for plotting relationships with the annual water balance, following the resampling method above.

### Gibbsite buffer

The pH of a solution exposed to calcite (CaCO3) and open to the atmosphere can be solved using an equation derived from the chemical equilibrium for CaCO3 (ref. 56):

\[
0 = H^+ (K_h/K_sK_h + PCO_2) + H^+(HK_sK_h - PCO_2 - K_hK_h - K_hPCO_2)
\]

where \(H^+\) is the hydrogen ion activity in moles, \(K_h\) is the solubility constant for CaCO3 (in units of \(\text{mol}^{-1}\)), \(K_s\) is the dissociation constant for water (in \(\text{mol}^{-1}\)), \(K_h\) and \(K_s\) are the first and second dissociation constants of carbonic acid (in \(\text{mol}^{-1}\)), \(K_h\) is Henry’s constant (in \(\text{mol}^{-1}\)), and \(PCO_2\) is the partial pressure of CO2 (in atm). We solved this equation for \(H^+\) at 25°C and a \(PCO_2\) of 3.45 × 10^-4 atm using the package rootSolve in R and published parameters. The \(PCO_2\) value of 3.45 × 10^-4 atm reflects \(PCO_2\) imposed by laboratory measurement conditions at standard atmospheric pressure, based on the ambient CO2 mole fraction in 1986, the median measurement date of the data. Older measurements made at lower atmospheric CO2 levels may reflect a slightly higher calcite equilibrium pH (that is, the expected pH is 8.3 before 1977). Because this difference is small and the majority of measurements were taken after this date, we report model fits for a single \(PCO_2\) value.

Calcite concentrations are approximate, and reported as CaCO3 equivalents. The NCSS database reports CaCO3 equivalents measured using a pressure calicimeter following acid dissolution, meaning that a range of carbonate minerals are included in the estimate. Also, because values are reported at a precision of 1%, some soils with <1% CaCO3 are probably reported with zero values, even if their pH reflects buffering by CaCO3.

### Global range of soil pH values

Data measured in water without a reported ratio were assumed to be measured at ratios of 1:1 or 1:2.5.


Extended Data Figure 1 | Soil pH versus calcite and exchangeable aluminium. Transparent points show a spatial sample of 20,000 measurements from the NCSS database. a, The relationship between soil pH at 0.5 m and CaCO$_3$ equivalents as a mass percentage. The yellow line shows the calculated pH of a solution in equilibrium with calcite and atmospheric CO$_2$ (345 parts per million) at 25 °C. b, The relationship between soil pH at 0.5 m and the log-ratio of exchangeable calcium (Ca$_X$) to exchangeable aluminium (Al$_X$), which is thought to control the pH of gibbsite-buffered soils. The yellow line is the fit by least-squares regression ($b_0 = 4.96$, $b_1 = 0.32$, $R^2 = 0.36$, $P < 0.01$).
Extended Data Figure 2 | Results of spatial resampling. Transparent points show a spatial sample of 20,000 measurements (a and b) and a random sample of 20,000 measurements (c and d). a, c, pH at 0.5 m depth versus MAP minus PET. b, d, The geographic distribution of measurements in the Americas.
Extended Data Figure 3 | Soil pH at 0.5 m depth versus alternative water-balance models. Transparent points show a spatial sample of 20,000 measurements of soil pH at 0.5 m depth. a, Soil pH versus MAP. b, Soil pH versus MAP minus PET estimated using the Priestley–Taylor method driven by CERES radiation data. c, MAP minus AET, from the LandFlux-EVAL synthesis. d, MAP minus PET estimated using the Priestley–Taylor method driven by GEWEX radiation data.
Extended Data Figure 4 | Soil pH at 0.1 m depth versus MAP minus PET. Transparent points show a spatial sample of 20,000 measurements of soil pH at 0.1 m depth. Side panels show histograms of MAP minus PET and soil pH, and yellow lines show predicted pH values of CaCO$_3$-buffered soils (8.2) and Al(OH)$_3$-buffered soils (5.1).
Extended Data Figure 5 | Calcite and exchangeable aluminium versus MAP minus PET. Transparent points represent a spatial sample of 20,000 measurements from the NCSS database. 

**a**, Calcite (CaCO$_3$) equivalents as mass percentage versus MAP minus PET. 

**b**, Exchangeable aluminium as a percentage of the effective cation exchange capacity versus MAP minus PET. These data are not reported for all samples in the NCSS database, and so points on the plot represent only the subset of the data with reported values.
Extended Data Figure 6 | Dry-climate soil pH versus seasonality, relief and carbonates. Transparent points show soil pH at 0.5 m depth in the driest quartile of MAP minus PET \((n = 5,000)\). a, Soil pH versus the coefficient of variation (CV) of precipitation. b, Soil pH versus local relief. c, Violin plots showing soil pH versus carbonate lithology. Panels d and e show the proportion of the observations with pH < 6.5, binned into deciles of the variable on the x axis; panel f shows the proportion in each lithologic category. Black lines show logistic regression fits, with associated chi-squared \((\chi^2)\) statistics and \(P\) values from likelihood ratio tests for precipitation CV \((\chi^2 = 167.65, \ P < 0.01)\), local relief \((\chi^2 = 76.5, \ P < 0.01)\) and carbonate lithology \((\chi^2 = 91.42, \ P < 0.01)\). Dashed lines show the proportion of observations with pH < 6.5.
Extended Data Figure 7 | Wet-climate soil pH versus seasonality, relief and carbonates. Transparent points show soil pH at 0.5 m depth in the wettest quartile of MAP minus PET (n = 5,000). a, Soil pH versus the coefficient of variation of precipitation. b, Soil pH versus local relief. c, Violin plots showing soil pH versus carbonate lithology. Panels d and e show the proportion of the observations with pH > 6.5, binned into deciles of the variable on the x axis; panel f shows the proportion in each lithologic category. Black lines show logistic regression fits, with associated χ² statistics and P values from likelihood ratio tests for precipitation CV (χ² = 3.5, P = 0.06), local relief (χ² = 61.29, P < 0.01) and carbonate lithology (χ² = 156.41, P < 0.01). Dashed lines show the proportion of observations with pH > 6.5.
## Extended Data Table 1 | Soil profile data sets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Provider</th>
<th>Reference</th>
<th># profiles used in analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>National Cooperative Soil Survey</td>
<td>United States Department of Agriculture</td>
<td>*</td>
<td>34,259 (0.1 m) 35,775 (0.5 m)</td>
</tr>
<tr>
<td></td>
<td>Natural Resources Conservation Service</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chinese National Soil Database</td>
<td>Chinese Soil Survey</td>
<td>*</td>
<td>2,183 (0.1 m) 2,370 (0.5 m)</td>
</tr>
<tr>
<td>World Inventory of Soil Emissions-Potentials</td>
<td>International Soil Reference and Information Center</td>
<td>45</td>
<td>2,505 (0.1 m) 2,682 (0.5 m)</td>
</tr>
<tr>
<td>Africa Soil Profile Database</td>
<td>International Soil Reference and Information Center</td>
<td>46</td>
<td>10,057 (0.1 m) 11,680 (0.5 m)</td>
</tr>
<tr>
<td>Australian National Soil Database</td>
<td>Commonwealth Scientific and Industrial Research Organization</td>
<td>*</td>
<td>4,389 (0.1 m) 5,959 (0.5 m)</td>
</tr>
<tr>
<td>Canadian National Soil Database</td>
<td>Canadian Soil Information Service</td>
<td>*</td>
<td>6,035 (0.1 m) 8,219 (0.5 m)</td>
</tr>
<tr>
<td>Soil Profile Analytical Database of Europe/Measured Parameters</td>
<td>European Soil Data Center</td>
<td>47</td>
<td>375 (0.1 m) 408 (0.5 m)</td>
</tr>
<tr>
<td>Brazilian National Soil Database</td>
<td>Luiz de Queiroz College of Agriculture</td>
<td>48</td>
<td>488 (0.1 m) 807 (0.5 m)</td>
</tr>
</tbody>
</table>

Data sets marked with an asterisk are publicly available on request from the data provider. The other datasets are described in refs 45-48.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Provider</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPCC Full Data Reanalysis v7.0</td>
<td>Global Precipitation Climatology Center</td>
<td>16</td>
</tr>
<tr>
<td>LandFlux-EVAL Synthesis</td>
<td>Institute for Atmospheric and Climate Science, ETH Zurich</td>
<td>18</td>
</tr>
<tr>
<td>GLIM Global Lithologic Map</td>
<td>Institute for Geology, Universität Hamburg</td>
<td>25</td>
</tr>
<tr>
<td>CERES Surface Radiation Budget v2.8</td>
<td>United States National Aeronautics and Space Administration</td>
<td>38</td>
</tr>
<tr>
<td>GEWEX Radiation Budget v3.0</td>
<td>United States National Aeronautics and Space Administration</td>
<td>39</td>
</tr>
<tr>
<td>CRU TS3.13 Global Surface Climatology</td>
<td>University of East Anglia Climate Research Unit</td>
<td>40</td>
</tr>
<tr>
<td>ETOPO1 Digital Elevation Model</td>
<td>United States National Centers for Environmental Information</td>
<td>41</td>
</tr>
<tr>
<td>MOD12 Land Cover Classes</td>
<td>United States National Aeronautics and Space Administration</td>
<td>42</td>
</tr>
<tr>
<td>GLASS Leaf Area Index</td>
<td>Global Land Cover Facility</td>
<td>43, 44</td>
</tr>
</tbody>
</table>

The models are described in refs 16, 18, 25, 38-44.