LETTER

doi:10.1038/nature20139

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

E.W. Slessarev[1], Y. Lin[2], N. L. Bingham[3], J. E. Johnson[4]†, Y. Dai[5], J. P. Schimel[1] & O. A. Chadwick[3]

Soil pH regulates the capacity of soils to store and supply nutrients, and thus contributes substantially to controlling productivity in terrestrial ecosystems[1]. 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[2,3]. 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[1]. 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[5,6]. 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 ($Al(OH)_3$)[7,8]. $CaCO_3$ 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[5]. 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 soil[2,3,8,9]. 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 leaches 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 demand[10], 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[4,8].

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 weathering[11], and thus interact with climate to influence soil pH over long timescales[8]. 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[9]. Alternatively, soils with short 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 validation[13,14]. However, these maps rely on spatial projections of soil taxonomy that are sometimes explicitly defined by climate[15], 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

[1]Department of Ecology, Evolution, and Marine Biology, University of California, Santa Barbara, California, USA. [2]Department of Environmental Science, Policy and Management, University of California, Berkeley, California, USA. [3]Department of Geography, University of California, Santa Barbara, California, USA. [4]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona, USA. [5]School of Atmospheric Sciences, Sun Yat-Sen University, Guangzhou, Guangdong, China. †Present address: Department of Global Ecology, Carnegie Institution for Science, Stanford, California, USA.

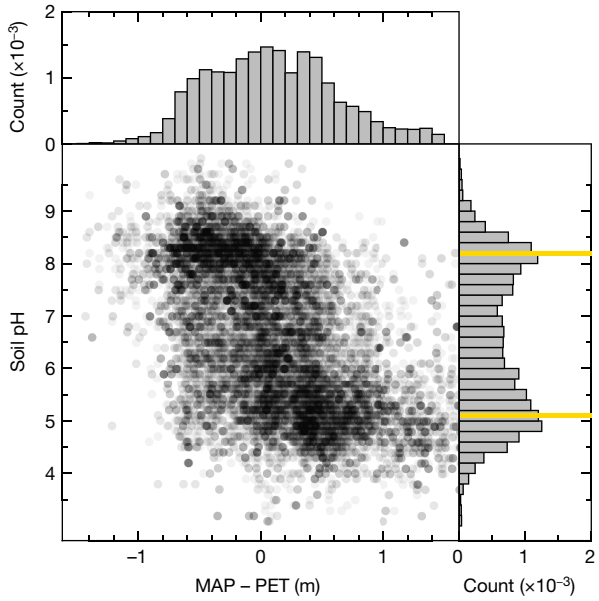


Figure 1 | Soil pH at 0.5 m depth versus annual water balance. Transparent points show a spatial sample of 20,000 measurements of soil pH at 0.5 m depth. Side panels show histograms of MAP minus PET and soil pH, and yellow lines show the predicted pH values of $CaCO_3$-buffered soils (8.2) and $Al(OH)_3$-buffered soils (5.1).

is a steep threshold between the two modes (Fig. 1). If we predict that leaching drives an immediate transition between $CaCO_3$ and $Al(OH)_3$ equilibria where MAP minus PET = 0, the predictions explain 42% (interquartile range 42%–45%) of the observed variation in pH (here 'variation' means the median absolute difference from the median; see Methods). The threshold pattern is robust, appearing when MAP is balanced with an ensemble estimate of actual evapotranspiration rates[18], or with simpler models of PET driven by different environmental data sets (Extended Data Fig. 3). Soil pH is not as strongly bimodal in surface soil—which may be affected by organic matter, biological cycling of Ca^{2+}, and agricultural liming—but the fundamental nonlinear pattern is still present (Extended Data Fig. 4). Furthermore, where MAP minus PET = 0, observed $CaCO_3$ concentrations diminish while exchangeable Al^{3+} increases, supporting the hypothesis that leaching drives a steep transition from carbonate to aluminium buffering at the global scale (Extended Data Fig. 5).

There are intriguing deviations from the central pattern—notably where acid soils occur in arid climates. These soils cluster into five regions: (1) the Sahel, (2) southern Africa, (3) northeastern Brazil, (4) Australia, and (5) mountains in western North America (Fig. 2). These acid soils may form where MAP minus PET is negative, but where appreciable leaching still occurs because of seasonal rainfall[19] or snowmelt[3]. Additionally, geologic constraints on Ca^{2+} supply may explain acid soils in regions (1) to (4), which are low-relief continental surfaces where erosion is probably limited, and where conditions were more humid during the Last Glacial Maximum[20–23]. Although the role of palaeoclimate in creating these acid soils is challenging to evaluate quantitatively, they are generally most common in both seasonal and low-relief environments (Extended Data Fig. 6). The prevalence of acid soils in arid, low-relief landscapes is consistent with the idea that depletion of Ca-bearing minerals might irreversibly constrain pH over long timescales, even in dry climates[9].

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[24], and might prevent soil acidification by sustaining Ca^{2+} supply to the soil solution. Comparison with a global lithologic map[25] shows that soil profiles in the wettest quartile of MAP minus PET are 2.6 times more likely to have a pH value > 6.5 when they fall within 1° grid cells containing carbonate bedrock (Extended Data Fig. 7). In regions (9) and (10), active volcanoes may produce easily weathered silicate minerals that could buffer pH outside the range of $Al(OH)_3$ equilibrium[26]. 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[27]. 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[8]. Not coincidentally, neutral-range soils are intensively cultivated, because they cluster in sub-humid climates with sufficient rainfall

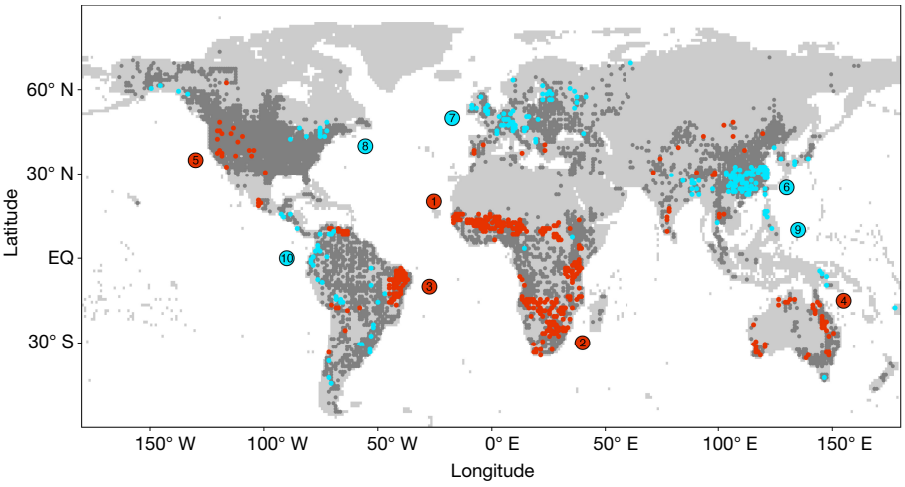


Figure 2 | Outliers from the global relationship between pH and MAP minus PET. Points show the centres of 1° grid cells containing soil profiles (n = 4,488). Cells in the driest quartile of MAP minus PET (in which the majority of profiles have pH < 6.5) are plotted in red, whereas cells in the wettest quartile of MAP minus PET (in which the majority of profiles have pH > 6.5) are plotted in blue, with the remaining cells shown in grey. Numbered regions are listed in the text.

for agriculture, but retain nutrients more effectively than acid soils[28]. 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.

1. Brady, N. C. & Weil, R. R. in *The Nature and Properties of Soils* 15th edn, 375–419 (Pearson Education, 2002).
2. Chadwick, O. A. *et al.* The impact of climate on the biogeochemical functioning of volcanic soils. *Chem. Geol.* **202,** 195–223 (2003).
3. Dahlgren, R. A., Boettinger, J. L., Huntington, G. L. & Amundson, R. G. Soil development along an elevational transect in the western Sierra Nevada, California. *Geoderma* **78,** 207–236 (1997).
4. Jenny, H. in *Factors of Soil Formation: A System of Quantitative Pedology* 104–196 (Dover, 1994).
5. Sposito, G. in *The Chemistry of Soils* 2nd edn, 275–295 (Oxford, 2008).
6. George, E., Horst, W. & Neumann, E. in *Mineral Nutrition of Higher Plants* (ed. Marschner, P.) 409–455 (Elsevier, 2012).
7. van Breemen, N., Mulder, J. & Driscoll, C. T. Acidification and alkalization of soils. *Plant Soil* **75,** 283–308 (1983).
8. Chadwick, O. A. & Chorover, J. The chemistry of pedogenic thresholds. *Geoderma* **100,** 321–353 (2001).
9. Vitousek, P. M. & Chadwick, O. A. Pedogenic thresholds and soil process domains in basalt-derived soils. *Ecosystems* **16,** 1379–1395 (2013).
10. Budyko, M. I. in *Climate and Life* 317–370 (Academic Press, 1974).
11. Hilley, G. E., Chamberlain, C. P., Moon, S., Porder, S. & Willet, S. D. Competition between erosion and reaction kinetics in controlling silicate-weathering rates. *Earth Planet. Sci. Lett.* **293,** 191–199 (2010).
12. Porder, S. *et al.* Linking geomorphology, weathering and cation availability in the Luquillo Mountains of Puerto Rico. *Geoderma* **249/250,** 100–110 (2015).
13. Batjes, N. H. Harmonised soil property values for broad-scale modelling (WISE30sec) with estimates of global soil carbon stocks. *Geoderma* **269,** 61–68 (2016).
14. Shangguan, W., Dai, Y., Duan, Q., Liu, B. & Yuan, H. A global soil data set for Earth System Modeling. *J. Adv. Model. Earth Syst.* **6,** 249–263 (2014).
15. FAO-UNESCO. *FAO Soil Map of the World* http://www.fao.org/soils-portal/soil-survey/soil-maps-and-databases/faounesco-soil-map-of-the-world/en/ (The Food and Agriculture Organization of the United Nations, 1998).
16. Schneider, U. *et al. GPCC Full Data Reanalysis Version 7.0 at 1.0°: Monthly Land-Surface Precipitation from Rain-Gauges built on GTS-based and Historic Data* http://doi.org/10.5676/DWD_GPCC/FD_M_V7_050 (2015).
17. Leuning, R., Zhang, Y. Q., Rajaud, A., Cleugh, H. & Tu, K. A simple surface conductance model to estimate regional evaporation using MODIS leaf area index and the Penman–Monteith equation. *Wat. Resour. Res.* **44,** 1–17 (2008).
18. Mueller, B. *et al.* Benchmark products for land evapotranspiration: LandFlux-EVAL multi-data set synthesis. *Hydrol. Earth Syst. Sci.* **17,** 3707–3720 (2013).
19. Milly, P. Climate, interseasonal storage of soil water, and the annual water balance. *Adv. Water Resour.* **17,** 19–24 (1994).
20. Kutzbach, J., Bonan, G., Foley, J. & Harrison, S. P. Vegetation and soil feedbacks on the response of the African monsoon to orbital forcing in the early to middle Holocene. *Nature* **384,** 623–626 (1996).
21. Fitzsimmons, K. E. *et al.* Late Quaternary paleoenvironmental change in the Australian drylands. *Quat. Sci. Rev.* **74,** 78–96 (2013).
22. Auler, A. S. & Smart, P. L. Late Quaternary paleoclimate in semiarid northeastern Brazil from U-series dating of travertine and water-table speleothems. *Quat. Res.* **55,** 159–167 (2001).
23. Chase, B. M. & Meadows, M. E. Late Quaternary dynamics of southern Africa's winter rainfall zone. *Earth Sci. Rev.* **84,** 103–138 (2007).
24. Ford, D. & Williams, P. in *Karst Hydrogeology and Geomorphology* 1–8 (Wiley, 2007).
25. Hartmann, J. & Moosdorf, N. The new global lithological map database GLiM: a representation of rock properties at the Earth surface. *Geochem. Geophys. Geosyst.* **13,** Q12004 (2012).
26. Nanzyo, M., Dahlgren, R. & Shoji, S. in *Volcanic Ash Soils: Genesis, Properties and Utilization* 145–188 (Elsevier, 1993).
27. Ulrich, B. Natural and anthropogenic components of soil acidification. *Z. Pflanzenernährung Bodenkunde.* **149,** 702–717 (1986).
28. Vitousek, P. M. *et al.* Soils, agriculture, and society in precontact Hawai'i. *Science* **304,** 1665–1669 (2004).

Acknowledgements We thank P. Vitousek, S. Fendorf, X. Feng and C. Kouba 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 Y.D. Statistical analysis and chemical calculations were performed by E.W.S. The manuscript was written by E.W.S. with input from all authors.

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.

METHODS

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, α (ref. 29). We also explored the relationship between soil pH and the difference of MAP and AET, which we represented using the mean of the diagnostic data included in the LandFlux-EVAL synthesis[18]. By definition, MAP minus AET cannot take negative values, but approaches zero where PET > MAP[10]. 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[17] partitions evaporation from the plant canopy (E_c) and soil (E_s). E_c is estimated using the Penman–Monteith equation[30], 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, α (ref. 31). Thus, we substituted the Priestley–Taylor model for the equilibrium model to represent soil evaporation in the Penman–Monteith–Leuning formula. The combined evaporation from canopy and soil are given by the equation:

$$\lambda E_{tot} = (sA_c + \rho c_p D_a G_a)/(s + \gamma(1 + G_a/G_c)) + \alpha(sA_s)/(s + \gamma) \quad (1)$$

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}), λ is the latent heat of vaporization of water (MJ kg^{-1}), s is the slope of the saturation vapour pressure curve (kPa °C^{-1}), ρ is the density of air (kg m^{-3}), c_p is the specific heat of air at constant pressure (MJ kg^{-1} °C^{-1}), D_a is the vapour pressure deficit (kPa), γ is the psychrometric constant (kPa °C^{-1}), G_a is the aerodynamic conductance (m d^{-1}), and G_c is the canopy conductance (m d^{-1}). Radiation is partitioned between canopy and soil by two equations[17]:

$$A_c = A_{tot}(1 - e^{-k_a L}) \quad (2)$$

$$A_s = A_{tot} - A_c \quad (3)$$

where A_{tot} is equal to R_n (the soil heat flux is assumed to be negligible), L is the leaf area index, and k_a is an extinction coefficient. Canopy conductance (G_c) is constrained by maximum stomatal conductance (g_{sx}), and modified by factors that represent dependence on light availability and vapour-pressure deficit[17]:

$$G_c = (g_{sx}/k_Q)\ln[(Q_h + Q_{50})/(Q_h \exp(-k_Q L) + Q_{50})][1/(1 + D_a/D_{50})] \quad (4)$$

where Q_h is photosynthetically active radiation at the top of the canopy (half of the incoming shortwave radiation), Q_{50} is a half-saturation constant for Q_h, D_{50} 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[32]. The parameters k_a and k_Q were both set equal to 0.6 m^{-1}, while Q_{50}, and D_{50} were set equal to 2.6 MJ m^{-2} d^{-1} and 0.8 kPa (ref. 32). The maximum stomatal conductance, g_{sx}, was set equal to 0.006 m s^{-1}, which is a reasonable mean estimate for natural vegetation[33], and scaled to a daily time step. The aerodynamic conductance, G_a, 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[32], 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 0.01 m s^{-1}. All other parameters were calculated or obtained from the Food and Agriculture Organization (FAO) guidelines[34].

Priestley–Taylor model. The Priestley–Taylor model for PET uses a single parameter, α, to account for adiabatic component of latent heat transfer[29]. While α may vary as a function of meteorological conditions[30], a standard α value of 1.26 has been applied successfully at large scales[35]. Priestley–Taylor PET is given by the equation:

$$E_{tot} = \alpha s A/(s + \gamma) \quad (5)$$

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 Full Data Reanalysis, Version 7.0[16]. MAP was calculated 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[36]. We corrected for systematic rain gauge measurement error using static monthly under-catch corrections[37] 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[38], and mean monthly values from the NASA/GEWEX surface radiation budget version 3.0 over the years 1984–2007[39]. 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[40]. Atmospheric pressure was obtained using mean elevations from the ETOPO1 global digital elevation model[41] in each 1° cell and correcting using the ideal gas law[34]. Land cover classes were obtained from the NASA MODIS satellite product MOD12[42] 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 set[43,44], 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[41]. Local relief was calculated as the difference between maximum and minimum elevations within a 10-km radius of each 1-arcminute cell centre. 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)[25]. 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[49]. 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[50]. 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 units[50], 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 node that contained profiles, and then randomly drawing 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 sampling distributions that appear 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, R^2) are inappropriate[51]. 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 $\mathrm{MAD_{residuals}}/\mathrm{MAD_{data}}$. This metric is analogous to R^2, 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[52] 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[53,54]. Likelihood ratio tests were used to compare regressions against the null hypothesis that the proportion of outliers is uniform with respect to each predictor[53]. 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 aluminium chemistry. We used the NCSS database to validate chemical calculations and determine the relationship between climate, calcite ($CaCO_3$), and exchangeable aluminium (Al_X). We used the NCSS database for this purpose because it contains a large number of measurements of $CaCO_3$ and Al_X using consistent methods[55], 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.

Calcite buffer. The pH of a solution exposed to calcite ($CaCO_3$) and open to the atmosphere can be solved using an equation derived from the chemical equilibria for $CaCO_3$ (ref. 56):

$$0 = H^4(2K_s/K_1K_2K_H \times p_{CO_2}) + H^3 - HK_wK_1K_H \times p_{CO_2} - K_1K_2K_H \times p_{CO_2} \quad (6)$$

where H is the hydrogen ion activity in moles, K_s is the solubility constant of $CaCO_3$ (in units of $\mathrm{mol^2\,l^{-2}}$), K_w is the dissociation constant for water (in $\mathrm{mol^2\,l^{-2}}$), K_1 and K_2 are the first and second dissociation constants of carbonic acid (in $\mathrm{mol\,l^{-1}}$), K_H is Henry's constant (in $\mathrm{mol\,l^{-1}\,atm^{-1}}$), and p_{CO_2} is the partial pressure of CO_2 (in atm). We solved this equation for H^+ at 25 °C and a p_{CO_2} of 3.45×10^{-4} atm using the package rootSolve[57] in *R* and published parameters[58,59]. The p_{CO_2} value of 3.45×10^{-4} atm reflects p_{CO_2} imposed by laboratory measurement conditions at standard atmospheric pressure, based on the ambient CO_2 mole fraction in 1985[60], the median measurement date of the data. Older measurements made at lower atmospheric CO_2 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 p_{CO_2} value.

Calcite concentrations are approximate, and reported as $CaCO_3$ equivalents. The NCSS database reports $CaCO_3$ equivalents measured using a pressure calcimeter following acid dissolution, meaning that a range of carbonate minerals are included in the estimate[55]. Also, because values are reported at a precision of 1%, some soils with $<1\%$ $CaCO_3$ are probably reported with zero values, even if their pH reflects buffering by $CaCO_3$.

Gibbsite buffer. The pH of a solution exposed to gibbsite ($Al(OH)_3$) in a soil with exchangeable aluminium (Al_X) depends on the ratio of Al_X to other exchangeable cations (Ca_X). In nature, the solubility of $Al(OH)_3$ and the exchange coefficients of clays do not follow the behaviour of purified laboratory solutions, and so the relationship between Al_X, Ca_X and pH must be estimated empirically[61]. To derive a typical pH for $Al(OH)_3$-buffered soils, we took the mean of all measurements from the spatial sample of the NCSS database with non-zero Al_X (pH = 5.1). Additionally, to validate the theoretical relationship between Ca_X/Al_X and pH, we fitted a model to measurements from the NCSS database taken at 0.5 m depth with non-zero Al_X and effective cation exchange capacity. The Gapon exchange model can be used to develop a log-linear relationship between Ca_X/Al_X and pH (ref. 61):

$$\mathrm{pH} = b_0 + b_1 \log_{10}(\mathrm{Ca_X/Al_X}) \quad (7)$$

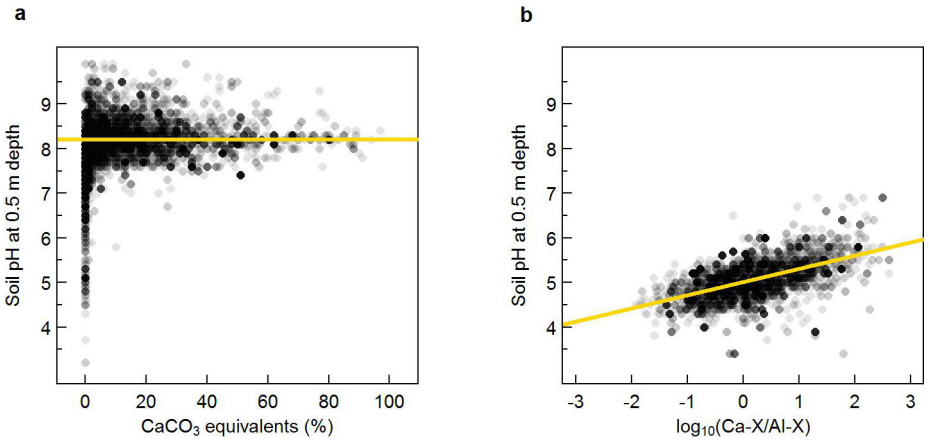
where b_0 and b_1 are fitted constants. To fit the model, we assumed that Al_X was equal to aluminium extractable in 1 M KCl, and Ca_X was equal to the effective cation exchange capacity minus Al_X. The data show a log-linear relationship between Ca_X/Al_X and pH (Extended Data Fig. 1, $b_0 = 4.96$, $b_1 = 0.32$, $R^2 = 0.36$, $P < 0.01$), supporting control of pH by Ca_X/Al_X. However, we note that the relationship appears slightly concave-curvilinear, suggesting that the Gapon model fails to account for the total activity of Al_X. This issue warrants further investigation.

Code availability. Code used to spatially resample soil profiles, calculate PET, and perform statistical analyses are maintained on GitHub and publicly archived online at http://dx.doi.org/10.5281/zenodo.61996. Code for pre-processing of raw data sets is available from the authors upon request.

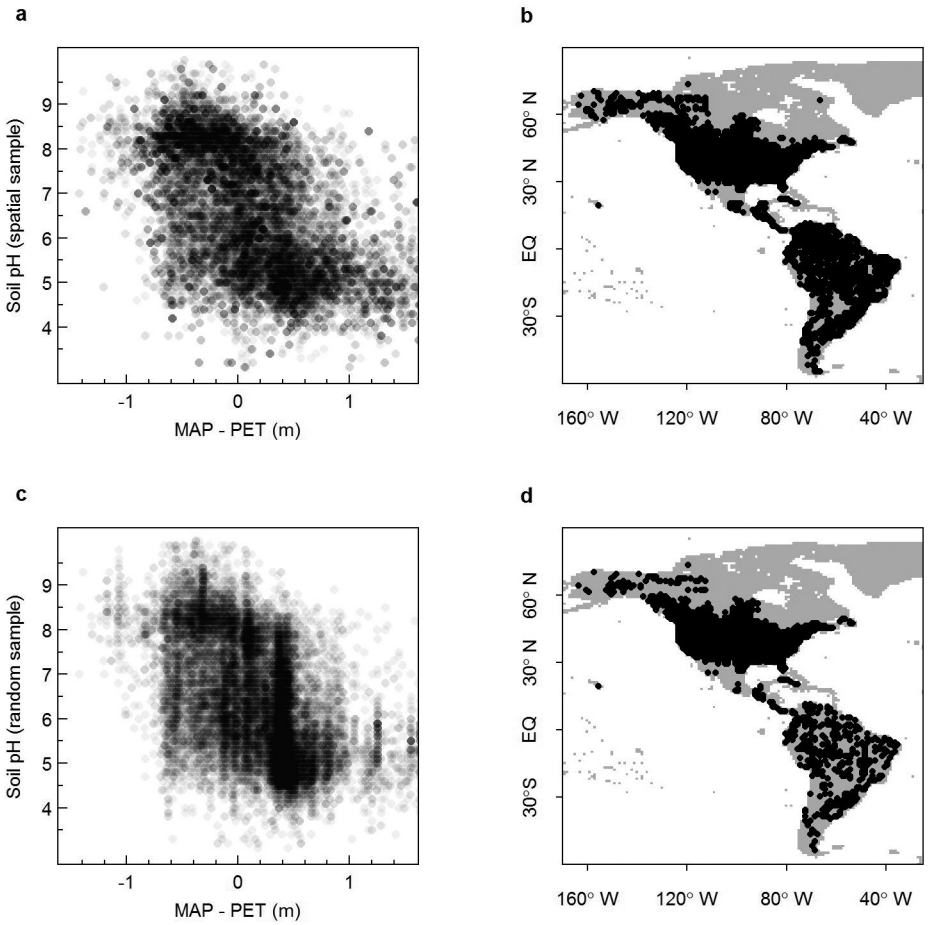
Data availability. All soil profile and meteorological data used in this study are publicly available from the sources listed in the text and in Extended Data Tables 1 and 2. Several of the soil profile databases are only available by direct request from the providing institutions (see Extended Data Table 1). As such, the combined soil profile data set used in this study is available from the authors upon request, given permission from providing institutions.

29. Priestley, C. H. B. & Taylor, R. J. On the assessment of surface heat flux and evaporation using large scale parameters. *Mon. Weath. Rev.* **100,** 81–92 (1972).
30. Monteith, J. & Unsworth, M. in *Principles of Environmental Physics* 4th edn, 217–247 (Elsevier, 2013).
31. Eichinger, W., Parlange, M. B. & Stricker, H. On the concept of equilibrium evaporation and the value of the Priestley-Taylor coefficient. *Wat. Resour. Res.* **32,** 161–164 (1996).
32. Zhang, Y. Q., Chiew, H. S., Zhang, L., Leuning, R. & Cleugh, H. A. Estimating catchment evaporation and runoff using MODIS leaf area index and the Penman Monteith equation. *Wat. Resour. Res.* **44,** W10420 (2008).
33. Kelliher, F. M., Leuning, R., Raupach, M. R. & Schulze, E. D. Maximum conductances for evaporation from global vegetation types. *Agric. For. Meteorol.* **73,** 1–16 (1995).
34. Allen, R. G., Pereira, L. S., Raes, D. & Smith, M. *Crop Evapotranspiration—Guidelines for Computing Crop Water Requirements*. FAO Irrigation and Drainage Paper 56, http://www.fao.org/docrep/x0490e/x0490e00.htm (FAO, 1998).
35. Fisher, J. B., Tu, K. P. & Baldocchi, D. D. Global estimates of the land–atmosphere water flux based on monthly AVHRR and ISLSCP-II data, validated at 16 FLUXNET sites. *Remote Sens. Environ.* **112,** 901–919 (2008).
36. Schneider, U., Becker, A. & Finger, P. GPCC's new land surface precipitation climatology based on quality-controlled in situ data and its role in quantifying the global water cycle. *Theor. Appl. Climatol.* **115,** 15–40 (2014).
37. Legates, D. R. & Willmott, C. J. Mean seasonal and spatial variability in gauge-corrected, global precipitation. *Int. J. Climatol.* **10,** 111–127 (1990).
38. Wielicki, B. A. *et al.* Clouds and the Earth's Radiant Energy System (CERES): an Earth observing system experiment. *Bull. Am. Meteorol. Soc.* **77,** 853–868 (1996).
39. Stackhouse, P. W. *et al.* The NASA/GEWEX surface radiation budget release 3.0: 24.5-Year Dataset. *GEWEX News* **21,** 10–12 (2011).
40. Harris, I., Jones, P. D., Osborne, T. J. & Lister, D. H. Updated high-resolution grids of monthly climatic observations—the CRU TS3.10 dataset. *Int. J. Climatol.* **34,** 623–642 (2014).
41. Amante, C. & Eakins, B. W. *ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis*. NOAA Technical Memorandum NESDIS NGDC-24, http://www.dx.doi.org/10.7289/V5C8276M (National Geophysical Data Center, NOAA, 2009).
42. Friedl, M. A. *et al.* Global land cover mapping from MODIS: algorithms and early results. *Remote Sens. Environ.* **83,** 287–302 (2002).
43. Liang, S. & Xiao, Z. *Global Land Surface Products: Leaf Area Index Product Data Collection (1985-2010).* http://www.dx.doi.org/10.6050/glass863.3004.db (Beijing Normal University, 2012).
44. Xiao, Z. *et al.* Use of general regression neural networks for generating the GLASS leaf area index product from time series MODIS surface reflectance. *IEEE Trans. Geosci. Remote Sens.* **52,** 209–223 (2013).

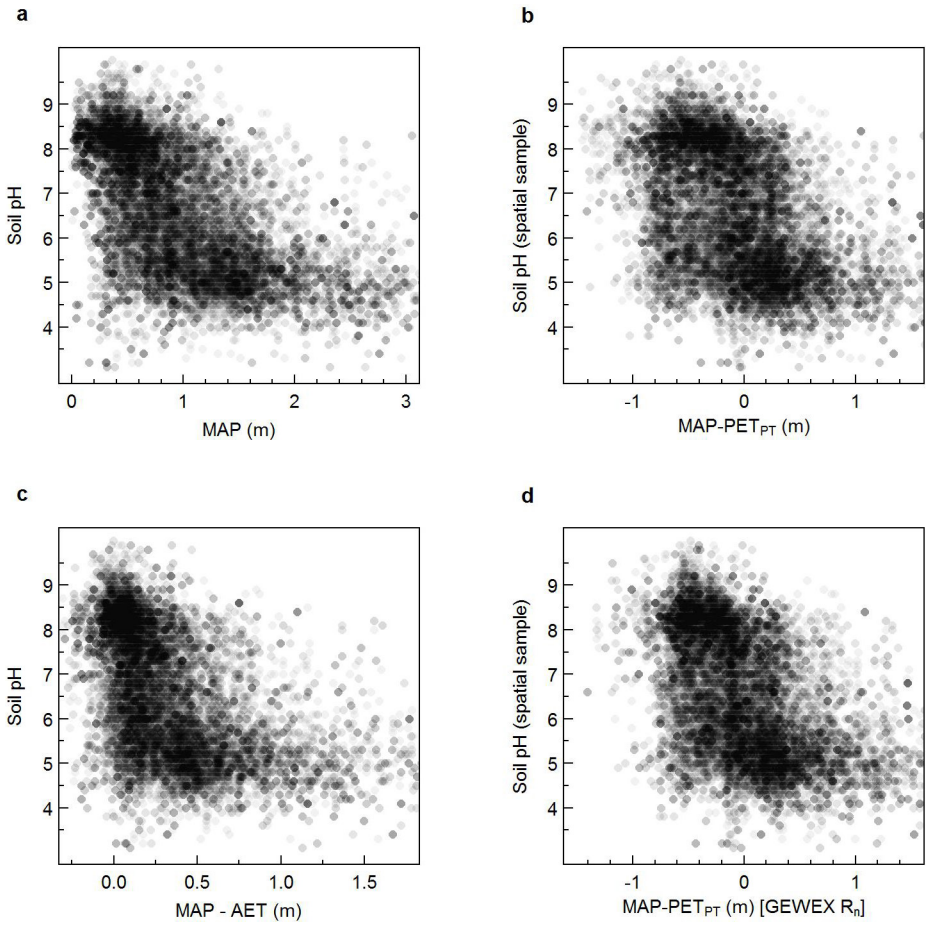
45. Batjes, N. H. Harmonized soil profile data for applications at global and continental scales: updates to the WISE database. *Soil Use Manage.* **25,** 124–127 (2009).
46. Leenaars, J. G. B., van Oostrum, A. M. J. & Ruiperez Gonzalez, M. *Africa Soil Profiles Database, Version 1.2. A Compilation of Georeferenced and Standardized Legacy Soil Profile Data for Sub-Saharan Africa (with Dataset).* ISRIC Report 2014/1 (Africa Soil Information Service (AfSIS) project, ISRIC, World Soil Information, 2012).
47. Hiederer, R., Jones, R. J. & Daroussin, J. Soil Profile Analytical Database for Europe (SPADE): reconstruction and validation of the measured data (SPADE/M). *Geograf. Tids. Dan. J. Geogr.* **106,** 71–85 (2006).
48. Cooper, M., Mendes, L. M. S., Silva, W. L. C. & Sparovek, G. A national soil profile database for Brazil available to international scientists. *Soil Sci. Soc. Am. J.* **69,** 649–652 (2005).
49. Jobbágy, E. G. & Jackson, R. B. The uplift of soil nutrients by plants: biogeochemical consequences across scales. *Ecology* **85,** 2380–2389 (2004).
50. Libohova, Z. *et al.* Converting pH 1:1 H_2O and 1:2 $CaCl_2$ to 1:5 H_2O to contribute to a harmonized global soil database. *Geoderma* **213,** 544–550 (2014).
51. Willmott, C. J., Matsuura, K. & Robeson, S. M. Ambiguities inherent in sums-of-squares-based error statistics. *Atmos. Environ.* **43,** 749–752 (2009).
52. Efron, B. Bootstrap methods: another look at the jackknife. *Ann. Stat.* **7,** 1–26 (1979).
53. Trexler, J. C. & Travis, J. Nontraditional regression analyses. *Ecology* **74,** 1629–1637 (1993).
54. R Core Team. *R: a Language and Environment for Statistical Computing.* https://www.R-project.org/ (Foundation for Statistical Computing, 2016).
55. Soil Survey Staff. In *Kellogg Soil Survey Laboratory Methods Manual Version 5.0 Soil Survey Investigations*. Report No 42 (eds Burt, R. & Soil Survey Staff) http://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/home/?cid=nrcs142p2_ (United States Department of Agriculture Natural Resources Conservation Service, 2014).
56. Plummer, N. L. & Busenberg, E. The solubilities of calcite, aragonite and vaterite in CO_2 H_2O solutions between 0 and 90°C, and an evaluation of the aqueous model for the system $CaCO_3$–CO_2–H_2O. *Geochim. Cosmochim. Acta* **46,** 1011–1040 (1982).
57. Soetaert, K. rootSolve: nonlinear root finding, equilibrium and steady-state analysis of ordinary differential equations. *R package version 1.6* https://cran.r-project.org/web/packages/rootSolve/index.html (2009).
58. Parkhurst, D. L. & Appelo, C. A. J. *Users Guide to PHREEC (Version 3.0): a Computer Program for Speciation, Batch-Reaction, One-Dimensional Transport, and Inverse Geochemical Calculations.* http://wwwbrr.cr.usgs.gov/projects/GWC_coupled/phreeqc/index.html (United States Geological Survey, 2013).
59. Sander, R. Compilation of Henry's law constants (version 4.0) for water as solvent. *Atmos. Chem. Phys.* **15,** 4399–4981 (2015).
60. Keeling, C. D. & Whorf, T. P. Atmospheric CO_2 concentrations derived from flask air samples at sites in the SIO network. In *Trends: A Compendium of Data on Global Change* http://cdiac.ornl.gov/trends/co2/sio-keel-flask/sio-keel-flaskmlo_c.html# (Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, United States Department of Energy, 2004).
61. Reuss, J. O., Walthall, P. M., Roswall, E. C. & Hopper, R. W. E. Aluminum solubility, calcium-aluminum exchange, and pH in acid forest soils. *Soil Sci. Soc. Am. J.* **54,** 374–380 (1990).



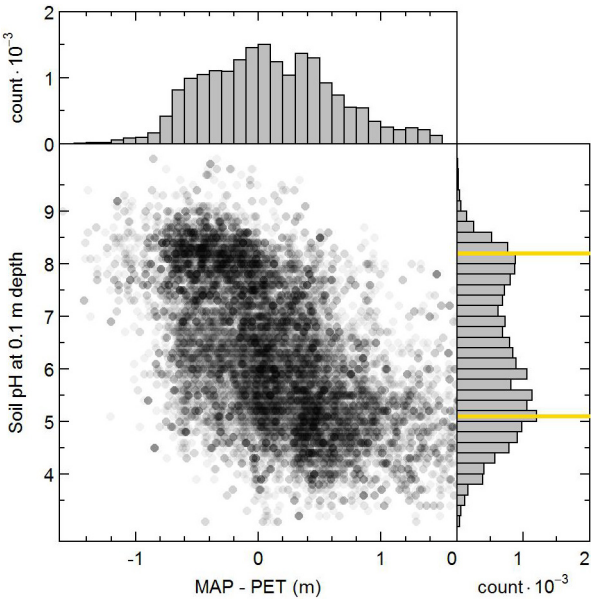
Extended Data Figure 1 | Soil pH versus calcite and exchangeable aluminium. Transparent points show a spatial sample of 20,000 measurements from the NCSS database. **a**, The relationship between soil pH at 0.5 m and $CaCO_3$ equivalents as a mass percentage. The yellow line shows the calculated pH of 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 ($b_0 = 4.96$, $b_1 = 0.32$, $R^2 = 0.36$, $P < 0.01$).



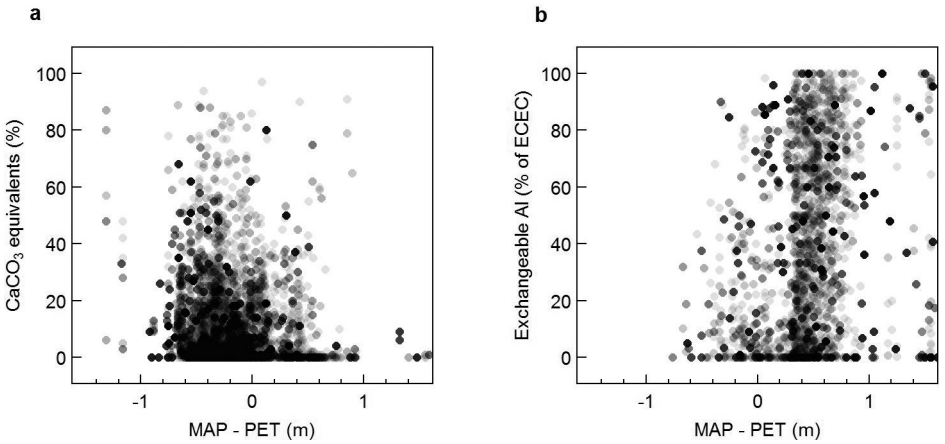
Extended Data Figure 2 | Results of spatial resampling. Transparent points show a spatial sample of 20,000 measurements (**a** and **b**) and a random sample of 20,000 measurements (**c** and **d**). **a**, **c**, pH at 0.5 m depth versus MAP minus PET. **b**, **d**, The geographic distribution of measurements in the Americas.



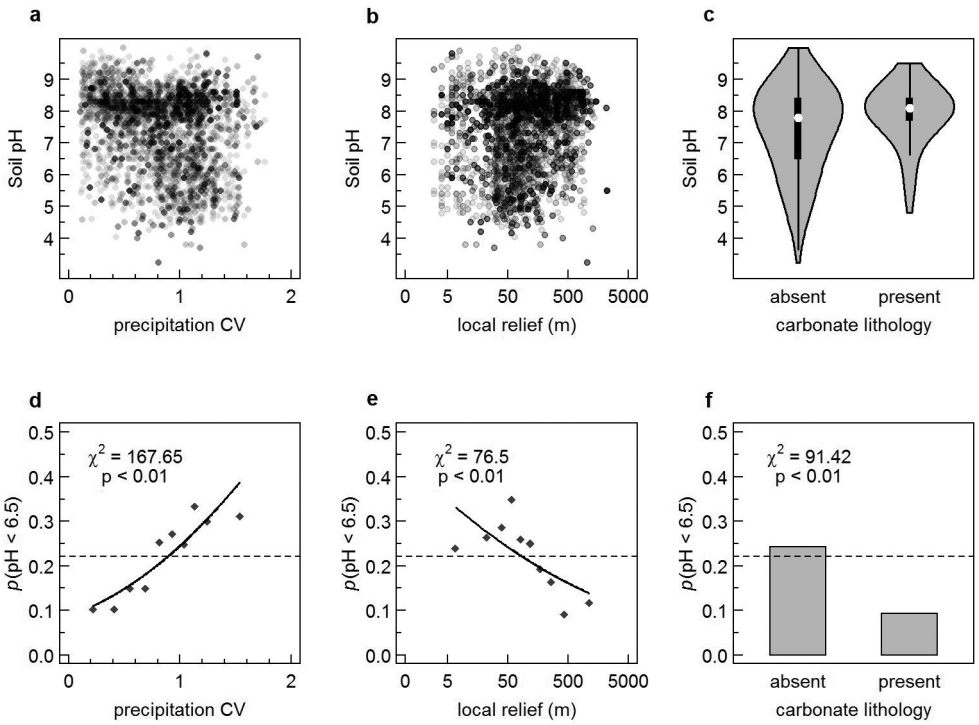
Extended Data Figure 3 | Soil pH at 0.5 m depth versus alternative water-balance models. Transparent points show a spatial sample of 20,000 measurements of soil pH at 0.5 m depth. **a**, Soil pH versus MAP. **b**, Soil pH versus MAP minus PET estimated using the Priestley–Taylor method driven by CERES 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.



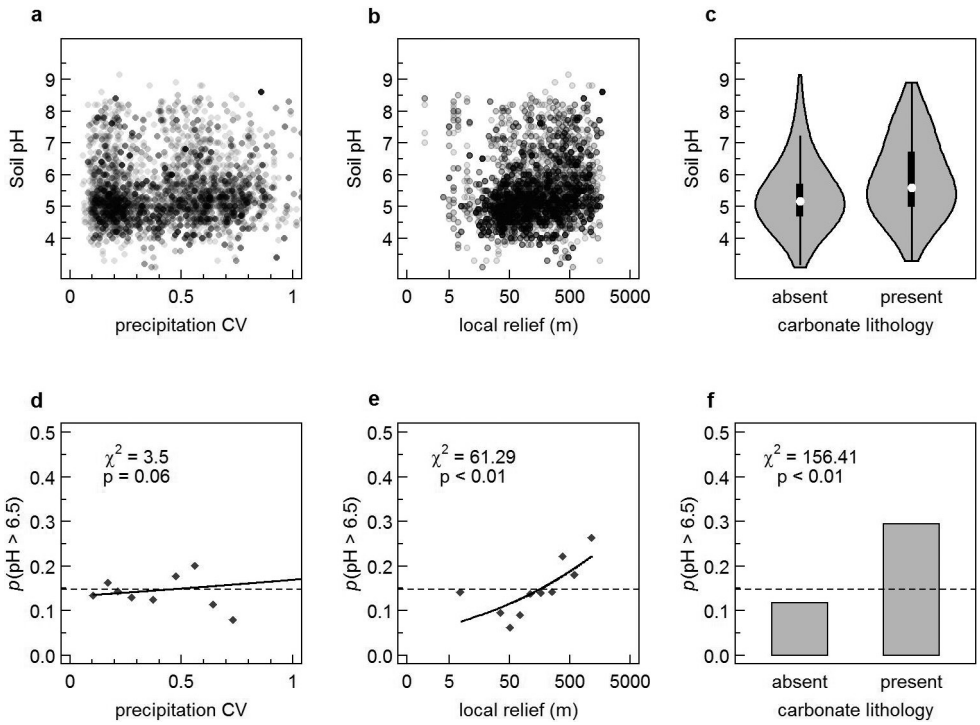
Extended Data Figure 4 | Soil pH at 0.1 m depth versus MAP minus PET. Transparent points show a spatial sample of 20,000 measurements of soil pH at 0.1 m depth. Side panels show histograms of MAP minus PET and soil pH, and yellow lines show predicted pH values of $CaCO_3$-buffered soils (8.2) and $Al(OH)_3$-buffered 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 (χ^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 χ^2 statistics and *P* values from likelihood ratio tests for precipitation CV ($\chi^2 = 3.5$, $P = 0.06$), local relief ($\chi^2 = 61.29$, $P < 0.01$) and carbonate lithology ($\chi^2 = 156.41$, $P < 0.01$). Dashed lines show the proportion of observations with pH > 6.5.

Extended Data Table 1 | Soil profile data sets

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

Data sets marked with an asterisk are publicly available on request from the data provider. The other datasets are described in refs 45–48.

Extended Data Table 2 | Gridded Environmental Data sets

Dataset	Provider	Reference
GPCC Full Data Reanalysis v7.0	Global Precipitation Climatology Center	16
LandFlux-EVAL Synthesis	Institute for Atmospheric and Climate Science, ETH Zurich	18
GLiM Global Lithologic Map	Institute for Geology, Universität Hamburg	25
CERES Surface Radiation Budget v2.8	United States National Aeronautics and Space Administration	38
GEWEX Radiation Budget v3.0	United States National Aeronautics and Space Administration	39
CRU TS3.13 Global Surface Climatology	University of East Anglia Climate Research Unit	40
ETOPO1 Digital Elevation Model	United States National Centers for Environmental Information	41
MOD12 Land Cover Classes	United States National Aeronautics and Space Administration	42
GLASS Leaf Area Index	Global Land Cover Facility	43, 44

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