Global patterns of relations between soil moisture and rainfall occurrence in ERA-40

Arien (H.A.) Lam

Department of Physical Geography, Utrecht University, Utrecht, Netherlands

Marc F.P. Bierkens

Department of Physical Geography, Utrecht University, Utrecht, Netherlands

Bart J.J.M. van den Hurk

Royal Netherlands Meteorological Institute, De Bilt, Netherlands

Short title: GLOBAL PATTERNS OF SOIL MOISTURE-RAINFALL OCCURRENCE RELATIONS

Abstract. We use a generalized linear model to statistically analyze the probability of daily precipitation occurrence, dependent on precipitation persistence, annual cycle, and soil moisture. By applying this method to the global ERA-40 re-analysis dataset, we reveil patterns of the precipitation occurrence variability as explained by each of the three variables and their interactions. These global patterns show: (1) known monsoon regions dominate the annual cycle component; (2) known regions of orographic uplifting dominate the persistence component; (3) the soil moisture component shows structure across all continents, but is most pronounced in the tropics and subtropics, and least pronounced in polar regions. In a surprisingly large part of the land surface, soil moisture influence on precipitation occurrence is of the same order of magnitude as the influence of the annual cycle.

Introduction

The dynamic role of terrestrial water in the hydrological cycle has been demonstrated in many regional and global climate studies. Especially, soil moisture plays a key role in land surface mass and energy balances, and relationships between soil moisture and rainfall affect regional climate and its variability.

Concepts of feedbacks between soil moisture and precipitation have been developed, that usually fall into one of the following two categories: feedbacks may consist of local recycling of present water [*Brubaker et al.*, 1993; *Eltahir and Bras*, 1996; *Trenberth*, 1999], or soil moisture may promote or hinder precipitation by altering boundary conditions [*Entekhabi et al.*, 1996; *Findell and Eltahir*, 1999; *Ek and Holtslag*, 2005; *Schär et al.*, 1999].

Validation of the role of soil moisture in both categories of concepts is problematic, both in modeling contexts and in situ [*Pan et al.*, 1996; *Betts*, 2004]. One of many issues hampering the straightforward analysis of relations between soil moisture and precipitation in the output of an Atmospheric General Circulation Model (AGCM) is the typical mixed, non-Gaussian probability distribution function of precipitation (with many zero's). Modeling precipitation characteristics is complex. Precipitation occurrence and precipitation amount are the most important characteristics that are mimicked by statistical precipitation models. These two variables are estimated either simultaneously [*Dunn*, 2004; *Durban and Glasbey*, 2001], or separately [*Srikanthan et al.*, 2005; *Harrold et al.*, 2003a, b]. For a review of statistical rainfall models, we refer to *Srikanthan and McMahon* [2001].

Analyses of relations between soil moisture and precipitation have mainly concentrated on precipitation amount. Among those, the GLACE project [*Koster et al.*, 2004, 2006; *Guo et al.*, 2006] is the most comprehensive. Other analyses are published in [*Koster and Suarez*, 1996; *Koster et al.*, 2000; *Dirmeyer*, 2000, 2005; *Savenije*, 1995]. Relations between soil moisture and rainfall probability—also known as precipitation occurrence or wet-day frequency—have rarely been investigated.

The purpose of our study is to identify relationships between soil moisture content and subsequent rainfall occurrence. We assume that while these relationships may exist throughout the world, the strength of the relations may vary as the underlying physical processes may vary spatially. Therefore, we focus on the strength of the relations, and their patterns. Global patterns of these relationships may then be compared with the patterns of land surface-atmosphere interaction strength as published by *Koster et al.* [2004, 2006], who focused their analysis on precipitation amount.

A generalized linear model (GLM) is used to statistically analyze the probability of daily precipitation occurrence, dependent on precipitation persistence, annual cycle, and soil moisture, and interactions between these variables. As the GLM is applied at each location of the ERA-40 dataset [*Uppala et al.*, 2005], patterns arise of relations between these three variables and precipitation occurrence. Analysis of global patterns offers an extension of validation opportunities, as spatial climate patterns have been well-studied for a long time. The precipitation persistence and the annual cycle in this context can be seen as nuisance variables, in the sense that each is an approximation for several global and local variables that act on the same timescale. The annual cycle variable, approximated by calendar months, represents sun intensity, sea surface temperature, snow cover, day length, surface temperature, among other variables. More precisely, it represents seasonal variability of all these variables together; interannual variability and variability within months is not accounted for. The persistence variable is primarily used to discern dry from wet episodes, and as such represents the presence of large-scale weather systems. It can also be seen as an autocorrelation parameter, accounting for system persistence and lagged cross-correlation of surface-atmosphere interactions. However, the patterns of the annual cycle and persistence offer a convenient benchmark since they can be visually validated: in the annual cycle patterns we expect to see monsoon regions, and the persistence patterns should show regions of orographic uplifting and regions where ocean winds enter the continents (the westerlies in the higher latitudes, both N and S hemisphere). If the resulting patterns resemble these expected patterns, we can have some confidence that the patterns of soil moistures' influence on precipitation occurrence are reasonable as well, provided that the statistical influence of the variables are of the same order of magnitude.

The rest of this paper is structured as follows. Section [ref to data] introduces the ERA-40 dataset and describes the data preparation. Section [ref to methods] explains the statistical methodology, including the meaning of test results in this context. Section [ref to results] presents and discusses results in the form of statistics and patterns. Finally, in section [ref to conclusions] we end with the conclusions.

Data

We used land surface water balance data from the ERA-40 global meteorological re-analysis, produced by the European Center for Medium-Range Weather Forecasts (ECMWF) [*Uppala et al.*, 2005]. The land surface parametrization of ERA-40 [*Viterbo and Beljaars*, 1995; *Van den Hurk et al.*, 2000] models the soil-atmosphere and soil-vegetation interactions and delivers a daily surface water and energy balance at each grid cell, during the entire period of 1957 to 2002. As differences in observation systems strongly influence biases in the hydrological cycle [*Hagemann et al.*, 2005; *Betts et al.*, 2005], we selected the satellite period (september 1978 - august 2002, 8766 days) to conduct our analysis. Grid cells with a land cover of less than 50% were excluded from the dataset, leaving 10407 grid cells for analysis.

Precipitation data from ERA-40 were retrieved and processed to obtain daily sums. To minimize spin-up biases, the 36-hour and 12 hour forecasts, started daily at 12 UTC, for four types of precipitation (convective snowfall, large-scale snowfall, convective rainfall, large-scale rainfall) were retrieved and subtracted from each other to obtain daily summed values over the 24-hour period from 00 to +24 hours UTC. Daily precipitation depth equals the sum of these four values. To transform the precipitation depth series into the precipitation occurrence series, a threshold of 0.1 mm/day was chosen to determine whether a day is wet or dry (following *New et al.* [1999, 2002]; *Buishand et al.* [2004]). A poor choice of threshold can possibly cause a lack of skill of the GLM, and also a trade-off might be expected with regard to the under-representation of sub-grid scale convective processes at the grid scale (threshold too high) and the over-representation of both numerical noise and biases induced by the model's constant drizzle (threshold too low). In order to assess the robustness of our method to the choice of precipitation threshold, several additional analyses were conducted, using thresholds from 0.05 mm/day up to 1.0 mm/day. The differences between the sets of results are very small, both in terms of proportion of explained deviance, and in terms of patterns. However, thresholds higher than 0.5 mm/day generate lower wet-day variability than thresholds lower than 0.3 mm/day, indicating that subgrid-scale convective processes are underrepresented with these higher thresholds. Consequently, in the subsequent sections we present only the results for the precipitation occurrence threshold of 0.1 mm/day.

Volumetric moisture content data were retrieved from the 36-hour forecasts (i.e. at +24 hours UTC). The volumetric moisture content of the upper meter of soil, obtained from the upper three layers in the ERA-40 dataset, forms the soil moisture variable used in this study. These are chosen because these three layers interact directly with vegetation in the ERA-40 land surface parametrization scheme [*Viterbo and Beljaars*, 1995; *Van den Hurk et al.*, 2000].

The time series data at each grid cell consists of 8766 records with four variables: precipitation occurrence (P_{bin}) as a binomial factor; precipitation occurrence on the previous day, used as an approximation of system persistence, as a binomial factor; annual cycle of seasonality, approximated by using calender months as a categorical factor; previous day volumetric soil moisture content as a continuous variable.

Methods

In this section, we describe the methodology of our analysis. We use previous day soil moisture and previous day precipitation occurrence to explain precipitation occurrence. This is done because in this way the obvious relationship of precipitation causing soil to get moist, is filtered out. This also implies that the results will only show a part of the influence of soil moisture on precipitation: processes that act within a day are not represented in our analysis.

All analyses were conducted in a widely used environment for statistical computing [*R Development Core Team*, 2006]. For estimation of each covariates' influence on the precipitation occurrence, a generalized linear model (GLM) with a logit link function

$$\text{logit}\, P_{bin,k} = \log\left(\frac{pr(P_{bin,k} = 1 \mid \mathbf{x}_k)}{1 - pr(P_{bin,k} = 1 \mid \mathbf{x}_k)}\right) = \mathbf{x}_k^T \beta \tag{1}$$

is used, where $\mathbf{x}_k$ is the data vector at time k (i.e. the value of all covariates, including interactions, and using dummy variables for the annual cycle), and β is the vector of corresponding coefficients (one for soil moisture, one for system persistence, eleven for the annual cycle, and 34 for interactions). In comparison with the classic linear regression model $pr(P_{bin,k} = 1 \mid \mathbf{x}_k) = \mathbf{x}_k^T \beta$, the logit function models the transition of probabilities in outcome of a dependent variable, given the change in the independent covariates. The outcome is confined to the domain of probabilities $[0, 1]$. Additivity of the effects in the right hand side of the equation is preserved, which allows analysis of variance, and consequently enables us to quantify the effects of the covariates on precipitation occurrence. *Grunwald and Jones* [2000] used a similar GLM within the context of Markov chain modeling of precipitation, and *Buishand et al.* [2004] used this approach in the context of statistical downscaling of AGCM output.

The coefficients in β are estimated by standard GLM methods [*McCullagh and Nelder*, 1989]. Rewriting equation 1 to obtain the estimated probability of precipitation occurrence is straightforward:

$$\begin{aligned} pr(P_{bin,k} \mid \mathbf{x_k}) &= \frac{\exp(\mathbf{x}_k^T \beta)}{1 + \exp(\mathbf{x}_k^T \beta)} \\ &= \frac{e^{\beta_0} \cdot e^{\beta_1 x_{1k}} \cdot e^{\cdots} \cdot e^{\beta_j x_{jk}}}{1 + e^{\beta_0} \cdot e^{\beta_1 x_{1k}} \cdot e^{\cdots} \cdot e^{\beta_j x_{jk}}} \end{aligned} \tag{2}$$

where x_{jk} is the value of covariate x_j at time k. As a measure of explained variation by the total model and each covariate, we use the deviance function, defined as

$$\mathrm{D} = 2\sum_{i=1}^{n} \left\{ y_i log \left(\frac{y_i}{\hat{\pi}_i} \right) + (m_i - y_i) \cdot \log \left(\frac{m_i - y_i}{m_i - \hat{\pi}_i} \right) \right\} \tag{3}$$

where n is the total number of categories i that emerge from the use of binary and categorical covariates (i.e. one category for each dummy variable) and m_i the number of days in each category (for instance all days in February), $\hat{\pi}_i$ the estimated rainfall occurrence and y_i the observed precipitation occurrence. Note that $\hat{\pi}_i = \sum_1^{m_i} \{pr(P_{bin,k} \mid \mathbf{x}_k \in i)\}$.

Deviance behaves similarly to sums of squares (R^2) in linear regression. Deviance attributed to the covariates is additive with respect to the total explained deviance, and can be interpreted similarly to explained variance in linear regression. We did not use the deviance as a goodness-of-fit statistic, but instead focussed on the components of explained deviance by their spatial patterns, and the links of these patterns to well-known phenomena.

As the common interpretation of rainfall occurrence probability as "wet-day frequency" presumes variation in wet and dry days, high proportions of explained deviance cannot be expected. Consider, for example, three consecutive days with estimated rainfall occurrence probability of 0.33, and one out of these days turns out to be "wet". This estimation, although "perfect" in the frequentist interpretation, errs at each of these days, and has a mean absolute error equal to 0.44. Another, partial explanation for the apparent lack of fit is the interannual variability of atmospheric processes and variables (for example sea surface temperature), which is not accounted for in our GLM.

We used methods of spatial data analysis and presentation as provided within the framework of statistical computing [*Pebesma and Bivand*, 2005].

Results and discussion

Figure 1.

– Figure 1 around here –

In this section we present the main results of our analyses. The scaled null deviances of precipitation occurrence (i.e. total variation of precipitation occurrence) at each grid point in the ERA-40 dataset are shown in Figure 1(top); proportions of the null deviance explained by the GLM at each grid point are shown in the bottom graph in Figure 1. The null deviance map is patterned with high values all around the world. Low variation in precipitation occurrence is expected and observed both in regions with lower precipitation occurrence (deserts, altiplanos), and in regions with high precipitation occurrence (tropical rain forests). Regions of low null deviance associated with low rainfall occurrence are the major deserts of Africa, Australia and Asia. Tropical Africa, Indonesia, the Amazon and the western coast of South-America are regions with low deviance associated with high rainfall occurrence.

The proportion of explained deviance by the GLM has a mean of 0.22. Apart from outliers in the desert regions (a maximum value of 0.87 in the western Sahara) the model generally explains up to 50 % of the null deviance. Regions of low explained deviance are all mid- and high-latitudes, the wet tropical regions, and the deserts. Regions with higher explained deviance are the Mexican west coast, the west coast of South-America, the southeastern fringe of the Amazon basin, most of the Sahel region, southern sub-tropical Africa, the Himalayas, Burma, and northern Australia.

– Figure 2 around here – **Figure 2.**

Figure 2 shows for all variables the deviance explained, both in absolute and in relative sense. The individual contributions of the variables are shown as well as the sum of all interactions between variables. At first glance, annual cycle and soil moisture stand out, while the influence of the interactions is negligible. The deviance explained by the annual cycle shows that the known monsoon regions stand out: the annual cycle in the Sahel, southern sub-tropical Africa and in eastern Asia influences the precipitation occurrence more than it does elsewhere in the world. In some very dry regions, where precipitation occurrence is exceptionally low, high values of proportions of null deviance explained by the annual cycle are found.

– Figure 3 around here – **Figure 3.**

Figure 2 also shows that persistence explains a relatively small part of the null-deviance, on average nine percent. To enhance the persistence pattern, we have plotted them at a separate scale in Figure 3. The figure shows a visible covariation with orography, which can be explained by orographic lifting of moist air that is (apparently) advected in separated multi-day episodes. Regions of strong covariance are not only the Andes, Rocky Mountains, Himalaya region, but also the Great Rift Valley in Africa. Also, temperate regions such as western Europe and Scandinavia show increased persistence. Here precipitation is expected throughout the year due to ocean winds (i.e. the westerlies) entering the continent. In some regions, for example the Arabian peninsula and central Australia, the apparent influence of persistence may be spurious, due to a small null deviance of precipitation occurrence.

The pattern of high influence of soil moisture in Figure 2 is similar to the pattern of high annual cycle influence. In the Sahel region, the band of soil moisture influence has its maximum more to the South compared to the pattern of the annual cycle's influence. Some of the "hot spots" in the patterns of soil moisture-precipitation feedback in our study are similar to the ensemble average coupling strength in the Global Land-Atmosphere Coupling Experiment [*Koster et al.*, 2004, 2006], such as the Sahel and Northern India. There are also differences such as central USA (not in our study) and the fringes of the Amazon (not in *Koster et al.* [2006]). However, there are several explanations for these differences. We are considering a different variable than *Koster et al.* [2006]—precipitation occurrence instead of amount. Also we are comparing one model approach (with re-analysis) with the average of 8 to 12 models by *Koster et al.*

[2006], where differences between models in the latter proved to be large as well. At present, we cannot say if our findings are consistent, complementary or contradictory to the results of *Koster et al.* [2006]. We also looked into the possibility (suggested by a reviewer) that our results may be dependent on mass balance deficits that exist in ERA-40. Deficits in the surface mass balance were added as independent variable in our GLM, and had near-zero explained deviance; they do not influence our results. Mass balance deficits in the atmospheric column were not assessed systematically. ERA-40 is not the only dataset that we could have used; NCEP/NCAR [*Kalnay et al.*, 1996] and JRA-25 [*Onogi et al.*, 2005] analyses—as well as many climate model output datasets—provide soil moisture and precipitation fields with daily values that are required for the GLM approach presented here. As several datasets are available to repeat our experiment, a possible next step is to assess the variability and commonality between the patterns resulting from these different data sources. Furthermore, a simultaneous analysis of land surface relations to both precipitation occurrence and amount—possibly by using Tweedie distributions, following *Dunn* [2004]—would shed more light on the issue of amount versus probability of precipitation.

– Figure 4 around here – **Figure 4.**

To assess spatial consistency of the GLM approach, three transects in the global dataset of GLM results are shown in Figure 4. Proportions of explained deviance are plotted against distance along the transects, and against soil moisture variance.

In the upper graph of Figure 4, showing the transect across the USA, the proportion of explained deviance of the three covariates express a marked difference between the western part and the eastern part of the transect. Soil moisture starts high in the west of the USA, maintaining a modest level at the Rocky Mountains, and gets very low in the mid-to-east part of the transect. Persistence, covarying with orography around the world, is strongest in the west, but still explains eight percent of the null deviance throughout in the eastern half of the transect. The influence of the annual cycle starts low, varies in the Rocky Mountains, stays low in the middle part of the transect and rises to eight percent at the east coast. Two factors seem to be responsible for these features. Firstly the dominant pattern of atmospheric circulation, bringing moist air from the west, interacting with a seasonal moist air pathway from the southeast and bending north and eastward over land. Secondly, the orography in the west (the Rocky Mountains) seems to restrict moisture availability in the rest of the transect, raising the threshold of rainfall event causation by local processes.

The middle graph shows the transect across the Sahara and Sahel. An extraordinary feature is the high proportion of deviance explained by the annual cycle from the Sahara to the dryer parts of the Sahel. The graph on the right shows the origin of these results: The highest scores of the annual cycle (circled in the right graph) are confined to regions where both soil moisture variance and variation in precipitation occurrence (see Figure 1) are low. This is an artifact of our formulation of the GLM, as the annual cycle variable in the GLM consists of eleven coefficients that are to be estimated, regardless how small the variability in precipitation occurrence is. The high values of relative explained deviance of annual cycle and soil moisture across the Sahel, lowering as the transect reaches the rain forest region of the Congo Basin, and constantly low persistence values indicate a strong land surface-atmosphere interaction in the Sahel.

In the Amazon transect, all three covariates show an active pattern. Explained proportions of deviance are low, as expected, at the equator (1000 km along the transect), and higher on both sides of the equator. A distinct low is observed around 2000 km along the transect. This pattern is consistent with the monsoonal circulation in South America [*Zhou and Lau*, 1998]. More southeast along the transect, the values climb as the transect enters the subtropical rain forests of the Brazilian Highlands, indicating a strong land surface-atmosphere interaction. Near the end of the transect, the lowering soil moisture and annual cycle values, accompanied by climbing persistence values indicate the proximity of the ocean and temperate eastern winds.

All three right-side graphs in Figure 4 show that the proportion of explained deviance by soil moisture is positively correlated to soil moisture variance, which is no surprise. However, in the bottom two graphs, the same goes for the proportion of explained deviance explained by the annual cycle, especially if the spurious results in the Sahara transect are discounted. This could indicate that our formulation of the GLM is not sufficiently able to distinguish between the influence of these two obviously covarying variables, thus overestimating the effect of one variable at the expense of the other's. However, in the transects there is little evidence (in the form of vertically symmetric curves) that this trade-off is spatially consistent.

- Figure 5 around here- **Figure 5.**

To explore the relative magnitude of explained deviances by the three variables we have plotted the logarithm of the ratio between the deviances in Figure 5. A value of 0 means that the two variables are equally important, 1 means that the explained deviance of the first variable is 10 times larger than the second, and a value of −1 the reverse. Figure 5 shows that the annual cycle has the largest influence on average, with strong dominace over other variables in the monsoon influenced climate zones: Sahel, South-subtropical Africa, N. Australia, Southeast Asia (see also Figure 2).

Compared to the other covariates, persistence explanation is more constant throughout the world. Persistence is dominant over soil moisture and annual cycle in the polar regions, and also in all regions where the total explained deviance of the model is low.

Soil moisture shows the smallest effects in both polar regions, where snow cover suppresses soil moisture variability and limits interaction between soil moisture and atmospheric processes. In the large deserts (Sahara, Australia) effects are also small because soil moisture hardly has natural variation. However, in other parts of the world the relative importance of soil moisture is considerable. The influence of soil moisture is in the same order of magnitude of the annual cycle's influence in 65 percent of the grid cells, which is a surprisingly large part of the land surface, and in 24 percent of the grid cells soil moisture influences precipitation occurrence stronger than the annual cycle. Soil moisture thus has a considerable influence on rainfall occurrence in large parts of the land surface.

Conclusions

Global patterns of relations between soil moisture and rainfall occurrence were examined. Soil moisture influences subsequent precipitation in a surprisingly large part of the land surface. This influence is of the same order of magnitude as the influence of either the annual cycle or precipitation persistence in a large part of the Earths' land surface. "Hot spots" of relative importance of soil moisture are the subtropical regions, notably the outer fringes of the monsoon influenced land surface: the southern Sahel, south Central Africa, the Amazon basin and northern India. The GLM in this paper does not generally explain much of the null deviance of precipitation occurrence in ERA-40. Further investigation may explain whether or not this is typical for the GLM approach used here, typical for the ERA-40 dataset, or typical for the chaotic nature of the hydroclimate.

Previous research has mainly focused on patterns of relationships between soil moisture and rainfall amount, whereas in this study precipitation occurrence was analyzed. At present, it is unknown to what extent the latter findings are consistent, complementary or contradictory to the former. Even taking into account that climate models are focused more on precipitation amount, it would be interesting to examine both occurrence and amount of precipitation simultaneously in a comparative study using multiple models.

Acknowledgment. This project was carried out in the framework of the Dutch NWO Program "Water". We thank two anonymous reviewers for helpful comments and suggestions, and Rens van Beek (UU), Jules Beersma (KNMI), Edzer Pebesma (UU) and Pier Sybesma (KNMI).

References

Betts, A., J. Ball, P. Viterbo, A. Dai, and J. Marengo, Hydrometeorology of the Amazon in ERA-40, *Journal of Hydrometeorology*, *6*, 764–774, 2005.

Betts, A. K., Understanding hydrometeorology using global models, *Bulletin of the American Meteorological Society*, *85*, 1673–1688, 2004.

Brubaker, K. L., D. Entekhabi, and P. S. Eagleson, Estimation of continental precipitation recycling, *Journal of Climate*, *6*, 1077–1089, 1993.

Buishand, T., M. Shabalova, and T. Brandsma, On the choice of the temporal aggragation level for statistical downscaling of precipitation, *Journal of Climate*, *17*, 1816–1827, 2004.

Dirmeyer, P. A., Using a global soil wetness dataset to improve seasonal climate simulation, *Journal of Climate*, *13*, 2900–2922, 2000.

Dirmeyer, P. A., The land surface contribution to the potential predictability of boreal summer season climate, *Journal of Hydrometeorology*, *6*, 618–632, 2005.

Dunn, P. K., Occurrence and quantity of precipitation can be modelled simultaneously, *International Journal of Climatology*, *24*, 1231–1239, 2004.

Durban, M., and C. A. Glasbey, Weather modelling using a multivariate latent Gaussian model, *Agricultural and Forest Meteorology*, *109*, 187–201, 2001.

Ek, M. B., and A. A. M. Holtslag, Evaluation of a land-surface scheme at Cabauw, *Theoretical and Applied Climatology*, *80*, 213–227, 2005.

Eltahir, E., and R. Bras, Precipitation recycling, *Reviews of Geophysics*, *34*, 367–378, 1996.

Entekhabi, D., I. Rodriguez-Iturbe, and F. Castelli, Mutual interaction of soil moisture state and atmospheric processes, *Journal of Hydrology*, *184*, 3–17, 1996.

Findell, K. L., and E. A. B. Eltahir, Analysis of the pathways relating soil moisture and subsequent rainfall in Illinois, *Journal of Geophysical Research D: Atmospheres*, *104*, 31,565–31,574, 1999.

Grunwald, G. K., and R. H. Jones, Markov models for time series with mixed distribution, *Environmetrics*, *11*, 327–339, 2000.

Guo, Z., et al., GLACE: The global Land-Atmosphere Coupling Experiment. Part II: Analysis, *Journal of Hydrometeorology*, *7*, 611–625, 2006.

Hagemann, S., K. Arpe, and L. Bengtsson, Validation of the hydrological cycle in ERA-40, *Tech. Rep. 24*, ECMWF, 2005.

Harrold, T. I., A. Sharma, and S. J. Sheather, A nonparametric model for stochastic generation of daily rainfall amounts, *Water Resources Research*, *39*, SWC81–SWC812, 2003a.

Harrold, T. I., A. Sharma, and S. J. Sheather, A nonparametric model for stochastic generation of daily rainfall occurrence, *Water Resources Research*, *39*, SWC101–SWC1011, 2003b.

Kalnay, E., et al., The NCEP/NCAR 40-year reanalysis project, *Bulletin of the American Meteorological Society*, *3*, 437–471, 1996.

Koster, R. D., and M. J. Suarez, The influence of land surface moisture retention on precipitation statistics, *Journal of Climate*, *9*, 2551–2567, 1996.

Koster, R. D., M. J. Suarez, and M. Heiser, Variance and predictability of precipitation at seasonal-to-interannual timescales, *Journal of Hydrometeorology*, *1*, 26–46, 2000.

Koster, R. D., et al., Regions of strong coupling between soil moisture and precipitation, *Science*, *305*, 1138–1140, 2004.

Koster, R. D., et al., GLACE: The Global Land-Atmosphere Coupling Experiment. Part I: Overview, *Journal of Hydrometeorology*, *7*, 570 – 610, 2006.

McCullagh, P., and J. A. Nelder, *Generalized Linear Models*, vol. 37 of *Monographs on statistics and applied probability*, second ed., Chapman & Hall, London, 1989.

New, M., M. Hulme, and P. Jones, Representing twentieth-century space-time climate variability. Part I: Development of a 1961-90 mean monthly terrestrial climatology, *Journal of Climate*, *12*, 829–856, 1999.

New, M., D. Lister, M. Hulme, and I. Makin, A high-resolution data set of surface climate over global land areas, *Climate Research*, *21*, 1–25, 2002.

Onogi, K., et al., JRA-25: Japanese 25-year re-analysis project - progress and status, *Quarterly Journal of the Royal Meteorological Society*, *131*, 3259–3268, 2005.

Pan, Z., E. Takle, M. Segal, and R. Turner, Influences of model parameterization schemes on the response of rainfall to soil moisture in the central united states, *Monthly Weather Review*, *124*, 1786–1802, 1996.

Pebesma, E. J., and R. S. Bivand, Classes and methods for spatial data in R, *R News*, *5*, 9–13, 2005.

R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2006.

Savenije, H. H. G., Does moisture feedback affect rainfall significantly?, *Physics and Chemistry of the Earth*, *20*, 507–513, 1995.

Schär, C., D. Lüthi, U. Beyerle, and E. Heise, The soil-precipitation feedback: A process study with a regional climate model, *Journal of Climate*, *12*, 722–741, 1999.

Srikanthan, R., and T. A. McMahon, Stochastic generation of annual, monthly and daily climate data: A review, *Hydrology and Earth System Sciences*, *5*, 653–670, 2001.

Srikanthan, R., T. I. Harrold, A. Sharma, and T. A. McMahon, Comparison of two approaches for generation of daily rainfall data, *Stochastic Environmental Research and Risk Assessment*, *19*, 215–226, 2005.

Trenberth, K. E., Atmospheric moisture recycling: Role of advection and local evaporation, *Journal of Climate*, *12*, 1368–1381, 1999.

Uppala, S. M., et al., The ERA-40 re-analysis, *Quarterly Journal of the Royal Meteorological Society*, *131*, 2961–3012, 2005.

Van den Hurk, B. J. J. M., P. Viterbo, A. C. M. Beljaars, and A. K. Betts, Offline validation of the ERA-40 surface scheme, *Tech. Rep. 295*, ECMWF, 2000.

Viterbo, P., and A. C. M. Beljaars, An improved land surface parameterization scheme in the ECMWF model and its validation, *Journal of Climate*, *8*, 2716–2748, 1995.

Zhou, J., and K.-M. Lau, Does a monsoon climate exist over South America?, *Journal of Climate*, *11*, 1020–1040, 1998.

H.A. Lam, Department of Physical Geography, Faculty of Geosciences, Utrecht University, P.O. Box 80.115, 3508 TC Utrecht, The Netherlands (a.lam@geo.uu.nl)

Received ______________________

Submitted to JGR D:Atmospheres

This manuscript was prepared with AGU's LaTeX macros v5, with the extension package 'AGU^{++}' by P. W. Daly, version 1.6b from 1999/08/19.

422 Figure Captions

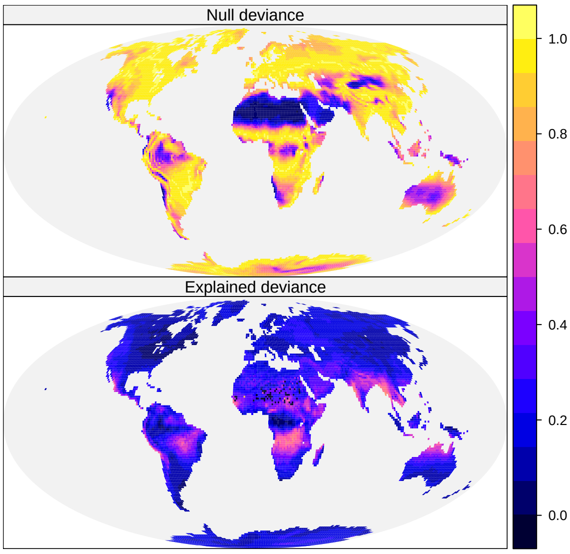

Figure 1. Patterns of null deviance of precipitation occurrence in the ERA-40 dataset (top) and patterns of the proportion of null deviance explained by the model (bottom). The null deviance values are scaled so that the global maximum value is 1, to allow a common color ramp for both graphs.

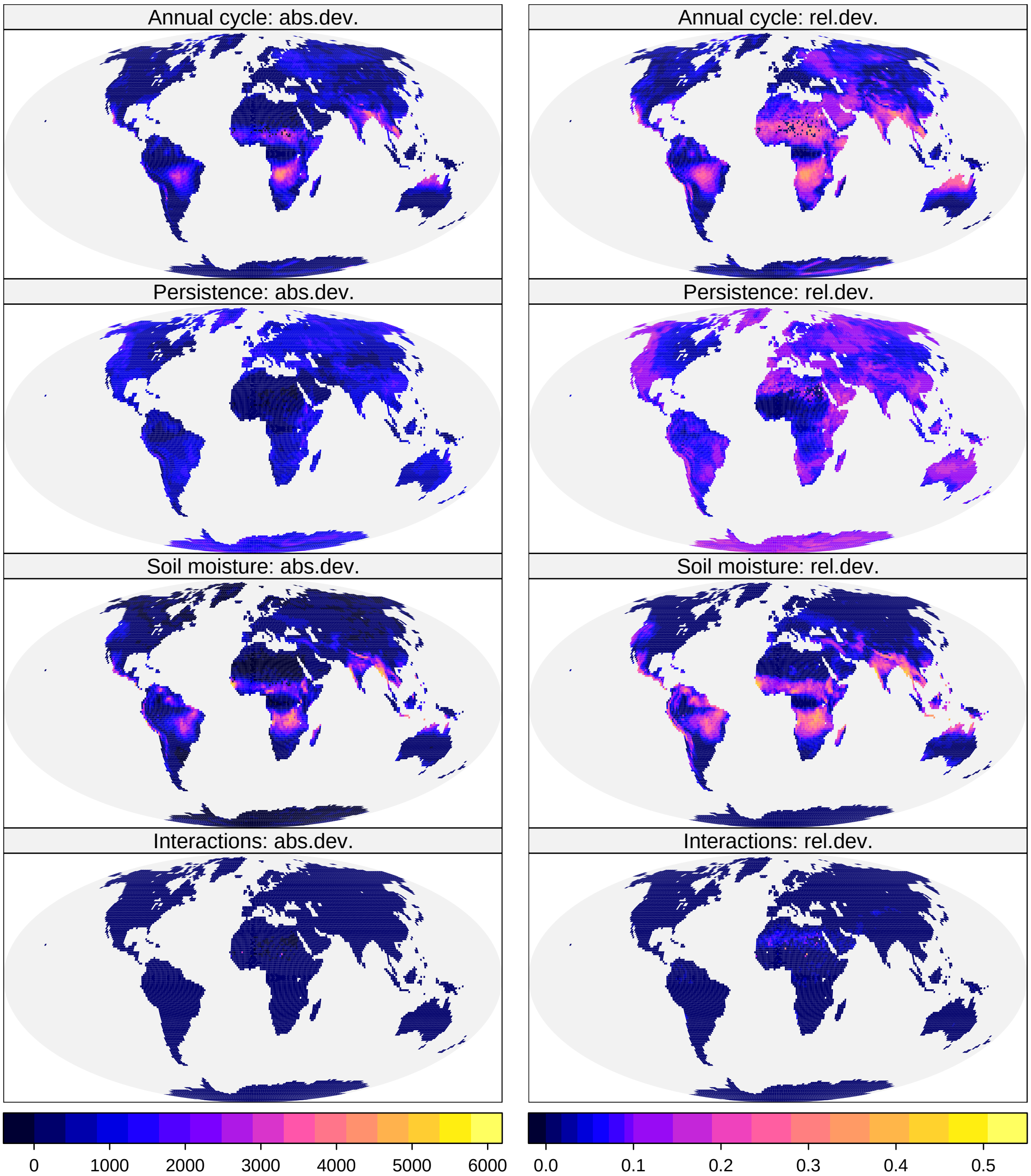

Figure 2. Deviance (left) and proportion of null deviance (right) explained by the components. Note that the color ramp on the right is stretched differently, to emphasize patterns in the moderate values. Deviance explained by interactions between the components of the GLM (bottom row) are aggregated values for all interactions between variables together.

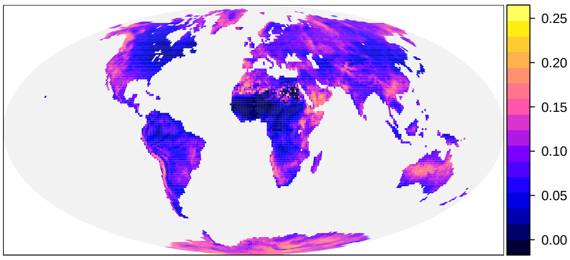

Figure 3. Proportion of deviance in ERA-40 explained by persistence.

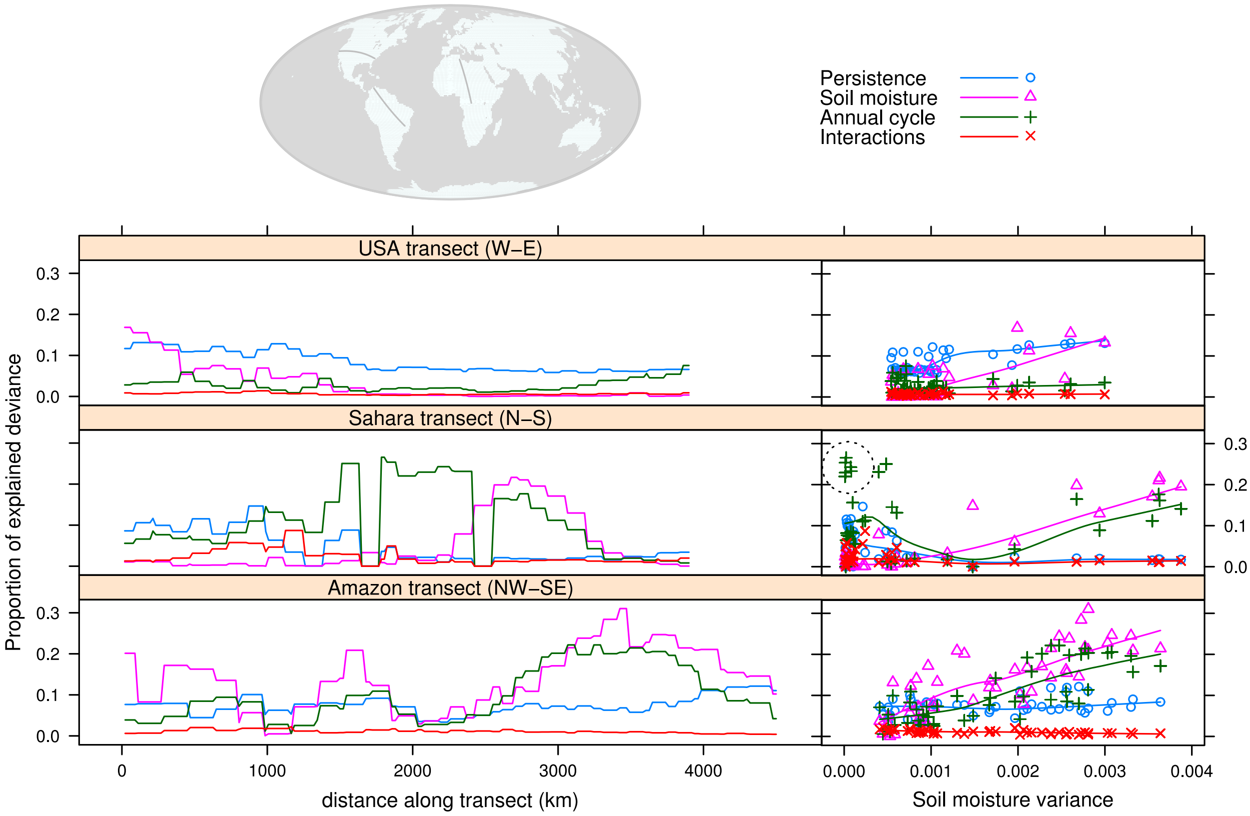

Figure 4. Transects of model results: the first transect crosses the USA from the west coast to east coast; the second transect crosses the Sahara and the Sahel, from Tunesia in the North to the Congo river basin in the South; the third crosses the Amazon basin, from Venezuela in the north west to the Brazilian Highlands in the southeast. Proportions of explained deviance are plotted against distance along the transect (on the left), and against soil moisture variance (on the right).

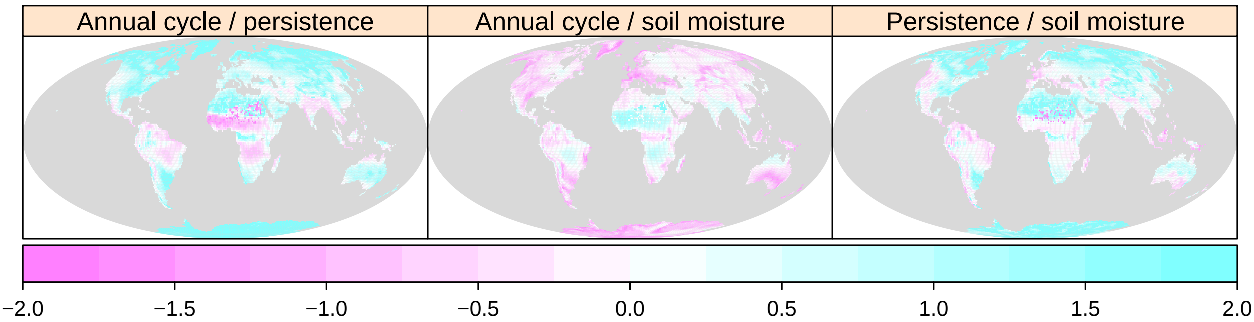

Figure 5. Relative order of magnitudes of each covariates' proportion of explained deviance compared to each of the other variables' proportion of explained deviance, computed by taking the logarithm of the fraction of each combination of covariates. Values outside the range [-2, 2] are clipped. For covariates to be in the same order of magnitude, the computed relative order of magnitude must be in the range [-1, 1].