NeuroImage 125 (2016) 556–570

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

A hemodynamic model for layered BOLD signals

Jakob Heinzle [a,*], Peter J. Koopmans [b], Hanneke E.M. den Ouden [c], Sudhir Raman [a], Klaas Enno Stephan [a,d,e]

[a] Translational Neuromodeling Unit, Institute for Biomedical Engineering, University of Zurich & ETH Zurich, Switzerland
[b] FMRIB Centre, Nuffield Department of Clinical Neurosciences, University of Oxford, UK
[c] Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands
[d] Wellcome Trust Centre for Neuroimaging, University College London, UK
[e] Max Planck Institute for Metabolism Research, Cologne, Germany

ARTICLE INFO

Article history:
Received 29 April 2015
Accepted 10 October 2015
Available online 17 October 2015

Keywords:
fMRI
Cortical layers
Dynamic causal modeling
Bayesian model comparison
Predictive coding

ABSTRACT

High-resolution blood oxygen level dependent (BOLD) functional magnetic resonance imaging (fMRI) at the sub-millimeter scale has become feasible with recent advances in MR technology. In principle, this would enable the study of layered cortical circuits, one of the fundaments of cortical computation. However, the spatial layout of cortical blood supply may become an important confound at such high resolution. In particular, venous blood draining back to the cortical surface perpendicularly to the layered structure is expected to influence the measured responses in different layers. Here, we present an extension of a hemodynamic model commonly used for analyzing fMRI data (in dynamic causal models or biophysical network models) that accounts for such blood draining effects by coupling local hemodynamics across layers. We illustrate the properties of the model and its inversion by a series of simulations and show that it successfully captures layered fMRI data obtained during a simple visual experiment. We conclude that for future studies of the dynamics of layered neuronal circuits with high-resolution fMRI, it will be pivotal to include effects of blood draining, particularly when trying to infer on the layer-specific connections in cortex — a theme of key relevance for brain disorders like schizophrenia and for theories of brain function such as predictive coding.

Introduction

Although only a few millimeters thin, the cerebral cortex is composed of microcircuits whose layered architecture plays a key role in cortical computation (Douglas and Martin, 2007; Heinzle et al., 2007; Bastos et al., 2012). Studying layer-specific computations non-invasively in humans would require two key ingredients: First, non-invasive high-resolution imaging to resolve cortical layers and, second, a modeling approach that explains the measured data as a function of neuronal interactions within and across layers.

Recent advances in high-field functional magnetic resonance imaging (fMRI) have made it feasible to measure blood oxygen level dependent (BOLD) signals from cortical structures with sub-millimeter resolution (Feinberg et al., 2010; Moeller et al., 2010; Poser et al., 2010; Heidemann et al., 2012). At this resolution both columnar architecture (Cheng et al., 2001; Yacoub et al., 2007; Yacoub et al., 2008) as well as cortical layers (Koopmans et al., 2010a; Polimeni et al., 2010; Koopmans et al., 2011; Olman et al., 2012) can be resolved. In rats, a highly specialized fMRI line-scanning method has been used to demonstrate that the layered pattern of temporal onsets of BOLD responses is in line with anatomical cortical connectivity (Yu et al., 2014).

Despite these encouraging technical developments, inferring on neural mechanisms at such high resolution with standard fMRI analysis procedures is complicated. This is due to fact that the strongest signals measured with fMRI depend on blood oxygenation mainly in venous compartments which are not evenly distributed over the cortical depth. Cortical blood supply is organized in a highly regular, layered fashion, similar to the neuroanatomical structure of cortex (Duvernoy et al., 1981; Weber et al., 2008). As illustrated in Fig. 1, arterial blood reaches the layers of cortex via diving arterioles that run perpendicular to cortex, passes the fine capillary bed within individual layers and flows back through ascending venules towards the pial surface (Duvernoy et al., 1981; Weber et al., 2008). This anatomical arrangement of blood flow has been modeled in detail (Boas et al., 2008; Reichold et al., 2009) and poses a fundamental problem for the analysis of layered BOLD activity since draining blood might affect the BOLD signal in lower (close to the white matter) and upper (close to the pial surface) layers differently. Standard hemodynamic models, like the "Balloon model" (Buxton et al., 1998) and subsequent extensions (Friston et al., 2000), assume that the measured BOLD response is driven

* Corresponding author.
E-mail address: heinzle@biomed.ee.ethz.ch (J. Heinzle).

http://dx.doi.org/10.1016/j.neuroimage.2015.10.025
1053-8119/

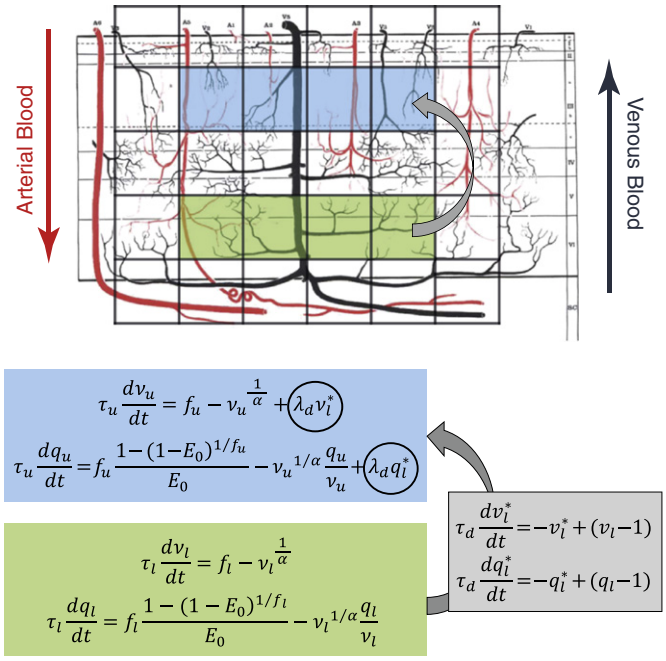

Fig. 1. Illustration of blood draining model equations. Top: Schematic visualization of fMRI voxel grid overlaid on an illustration of the layered architecture of blood flow (Duvernoy et al., 1981; reproduced with permission). Arterioles (red) and venules (dark blue) travel perpendicular to the cortical layers. Arrows indicate the directions of arterial and venous blood flow. Shaded voxels correspond to upper (blue) and lower (green) regions of interest whose BOLD signals interact through blood draining effects (gray arrow). Bottom: The local balloon model equations and the newly introduced blood draining (gray) effects.

by changes in relative blood volume and deoxyhemoglobin concentration in the venous blood. However, a fundamental assumption of this model is that the BOLD signal only depends on local neuronal activity. While this assumption seems adequate for conventional fMRI data analyses, it is problematic for high-resolution laminar fMRI since venous blood from deeper layers contributes to the BOLD signal in upper layers as it flows towards the pial surface.

In this work, we address this problem by introducing a novel extension to an established hemodynamic model (Buxton et al., 1998; Friston et al., 2000; Stephan et al., 2007). This extended model takes into account effects of cortical blood flow across layers and considers three different contributions to layer-wise BOLD measures: layer-specific neuronal inputs (e.g., synaptic inputs from remote regions), local neuronal connections across layers, and local blood flow effects across layers. To this end, the model incorporates distinct representations of neuronal connectivity across layers and a phenomenological description of venous blood flow effects perpendicular to the cortical surface; the latter allows BOLD activity in lower layers to contribute to the measured BOLD signal in upper layers via blood inflow, referred to as "blood draining" (BD) for the remainder of the paper. In order to evaluate the utility of this layered hemodynamic model, we use Bayesian model inversion and selection, implemented within the framework of dynamic causal modeling (DCM; Friston et al., 2003).

To prevent any misunderstandings, we would like to emphasize that this study does not present a model of layered BOLD measurements which strives for complete interpretability in physical and physiological terms, as recent models of non-layered BOLD (e.g., Havlicek et al., 2015). Notably, this study pursues a more modest ambition: it introduces a phenomenological description of blood draining effects across layers and examines (i) to what degree this relatively simple model can capture main features of layer-wise BOLD and (ii) the feasibility of model inversion, i.e., how well model parameters and structure can be identified from empirical data. This represents a first step towards establishing a hemodynamic component for models of effective connectivity which can operate on layer-wise BOLD data. Such future models are needed to test predictions from influential theories like predictive coding (Rao and Ballard, 1999; Friston, 2005) which postulate that supragranular and infragranular cortical layers convey different signals via their connections. It is possible that more ambitious and sophisticated biophysical models of blood flow across layers will be beneficial for this endeavor; however, this is an empirical question which will have to be adjudicated by model comparison in future work (see Discussion section).

Following the theoretical derivation of the model, we test the face validity and performance of the proposed model using both simulations and empirical analyses. First, we tested whether adding blood draining from lower to upper layers would reproduce key features of layer-specific BOLD signals as obtained from high-resolution fMRI in humans (e.g. in Siero et al., 2011). Second, we examined whether the model was capable of distinguishing between effects of neuronal connectivity and blood draining across layers, respectively. Third, we illustrate what effect the inclusion or exclusion of blood draining has on inferring the layered input structure from fMRI data. Fourth, we asked how well the parameters of the model could be inferred from simulated data where "ground truth" is known. Finally, we applied the model to fMRI data from a visual paradigm. Here, we used Bayesian model selection (BMS) to investigate which model provided a more convincing explanation for the observed data — the proposed model with blood draining effects across layers or an alternative model that allowed for between-layer differences in local hemodynamics.

Methods

In the following, we describe our novel model for layered hemodynamic responses. Starting from the standard hemodynamic model in DCM of fMRI — an extension of the Balloon model by Buxton et al. (1998) — we outline in detail the assumptions made in order to introduce hemodynamic coupling, from lower to upper layers.

The standard hemodynamic model in DCM

The standard hemodynamic model in DCM has been described in detail in previous work (Stephan et al., 2007). In this model, neurovascular coupling equations relate local changes in blood flow f to local changes in the neuronal activity x:

$$\begin{aligned}\frac{ds}{dt} &= x - \kappa s - \gamma(f-1)\\ \frac{df}{dt} &= s,\end{aligned} \tag{1}$$

where s is a vasodilatory signal, and f represents blood flow. x denotes the time course of neural activity, κ is the rate constant of the vasodilatory signal decay and γ specifies the rate constant for autoregulatory feedback by blood flow. These (and all further) hemodynamic states below are time-dependent and normalized to their values at rest. The changes in blood flow lead to local changes in relative blood volume v and in q, deoxygenated hemoglobin (deoxyHB) content of the venous blood. The dynamics of these two quantities are modeled using the Balloon model of Buxton et al. (1998).

$$\begin{aligned}\tau\frac{dv}{dt} &= f - v^{\frac{1}{\alpha}}\\ \tau\frac{dq}{dt} &= f\frac{1-(1-E_0)^{1/f}}{E_0} - v^{1/\alpha}\frac{q}{v}.\end{aligned} \tag{2}$$

Finally, the relative BOLD signal change is given by the nonlinear signal equation (Stephan et al., 2007)

$$\frac{\Delta S}{S} \approx V_0\left[k_1(1-q) + k_2\left(1-\frac{q}{v}\right) + k_3(1-v)\right], \tag{3}$$

where the coefficients k_i are field- and acquisition-dependent. In this paper, we restrict ourselves to simulations and data acquired at 3 T. A detailed description of the parameters k_i and their dependency on field strength is provided in the Appendix A.

In this work, we couple this hemodynamic model to the standard neural model used in DCM in order to test whether the model can disambiguate hemodynamic (blood flow) and neuronal (connectivity) influences across layers. Concretely, neuronal population activity x in Eq. (1) is provided by a simple linear form of the neuronal state equation in DCM where bilinear or nonlinear terms are omitted:

$$\frac{dx}{dt} = Ax + Cu. \tag{4}$$

Here, A is the static (fixed) connectivity between neuronal populations, and C describes the weights of driving inputs (experimentally controlled perturbations, e.g., sensory stimuli).

Incorporating blood draining effects in layered DCM

A critical assumption made by the hemodynamic model described above is that the hemodynamic response depends only on local neuronal activity. However, this assumption is violated when modeling layered gradient-echo BOLD responses, where layer-crossing draining veins have to be taken into account. Here, the changes required to include effects of intra-cortical blood draining in the hemodynamic model are presented for the simple case of two layered compartments: an upper or supragranular layer (close to the pial surface) and a lower or infragranular layer (close to the gray/white matter boundary). While this two-layer model is sufficient to investigate the importance of blood draining effects in layered fMRI data (and is adapted to the present resolution of layered human fMRI), it is straightforward to extend the model to more than two layers.[1]

Here, we propose a phenomenological description that captures two general features of how hemodynamic changes in lower layers should propagate towards the pial surface and, hence, influence the responses in upper layers: first, measurements of the blood flow velocity in small arterioles provide values on the order of millimeters per second (Santisakultarm et al., 2012), suggesting a delay of approximately one second between lower and upper layers. Second, due to dilution, the purely blood draining dependent signal part in upper layers should be smaller than the original signal in the lower layer. This decay can be described by two differential equations:

$$\begin{aligned}\tau_d \frac{dv_l^*}{dt} &= -v_l^* + (v_l - 1)\\ \tau_d \frac{dq_l^*}{dt} &= -q_l^* + (q_l - 1).\end{aligned} \tag{5}$$

Asterisks denote delayed volume and deoxy-HB concentration variables, respectively. Incorporating these effects into the equations of the classical Balloon model (Buxton et al., 1998) and introducing directed blood draining from lower to upper layers (Fig. 1) yields the following equations for the relative blood volume and deoxyHB concentration in the two layers (l: lower; u: upper):

$$\begin{aligned}\tau_l \frac{dv_l}{dt} &= f_l - v_l^{\frac{1}{\alpha}}\\ \tau_l \frac{dq_l}{dt} &= f_l \frac{1-(1-E_0)^{1/f_l}}{E_0} - v_l^{1/\alpha}\frac{q_l}{v_l}\\ \tau_u \frac{dv_u}{dt} &= f_u - v_u^{\frac{1}{\alpha}} + \lambda_d v_l^*\\ \tau_u \frac{dq_u}{dt} &= f_u \frac{1-(1-E_0)^{1/f_u}}{E_0} - v_u^{1/\alpha}\frac{q_u}{v_u} + \lambda_d q_l^*.\end{aligned} \tag{6}$$

Here, we have introduced two new parameters: The draining time constant τ_d controlling the delay and the coupling parameter λ_d representing the strength of the blood draining effect from the lower to the upper layer. Fig. 1 illustrates these equations. Please note that both parameters are restricted to positive values and that for $\lambda_d = 0$ the hemodynamic equations are exactly the same as for two independent cortical areas in the standard DCM (Stephan et al., 2007). Please note that the model described here is not a detailed physiological model of blood flow in vessels. Instead, it tries to capture qualitatively how draining blood from lower layers will affect the upper layer's BOLD response. Hence, there is no obvious direct physiological correspondence to the parameter λ_d. However, its effect can be summarized as the degree to which relative changes of blood volume as well as deoxyHB concentration in the lower layer will affect the corresponding values in the upper layer, respectively. See Fig. 2 for simulations that illustrate the role of the two newly introduced parameters.

Simulation of data

Simulations were conducted in order to address the following four questions. First, we simulated BOLD signal traces in order to verify whether the modeled responses were in agreement with empirical measurements of layered BOLD activity (Siero et al., 2011). Second, we asked, whether it was possible to distinguish models with different types of interactions across cortical layers. In particular, we compared models with blood draining to models with a neuronal connection between lower and upper layers. For this purpose, we generated data from models with and without across-layer blood draining, inverted the models for all generated data sets and finally employed model comparison (Penny et al., 2004; Penny et al., 2010; Penny, 2012) to infer which model had generated the data. Third, we investigated how well the structure of inputs to a layered circuit could be recovered from fMRI data generated with the model. Fourth, we tested how well known parameters of a generating model could be recovered from noisy synthetic BOLD time courses.

All simulations were based on an event-related scenario, where the driving input to the two layers consisted of 90 short events of 0.5 seconds duration. A third of these events drove both layers simultaneously, while the other two thirds targeted the individual layers (one third per layer). The time between events varied randomly between 3 and 10 s, leading to an "experimental" duration of less than 10 min.

The simulation parameters for the generating model (para_{gen}) are given in Table 1 (hemodynamic model) and in Supplementary Table 1 (neuronal model). In our simulations, we varied the signal to noise ratio, defined as the ratio of the standard deviation of the signal and the standard deviation of the noise ($\text{SNR} = \sigma_{\text{signal}} / \sigma_{\text{noise}}$). This definition of SNR is typically used in DCM and offers an intuitive measure of the ratio of the variability of the signal and noise. It compares the task induced signal with the noise of the data and is thus closely related to the contrast to noise ratio often used in fMRI as well as to the percentage

[1] For example a three layered network with infra-granular, granular, and supragranular neuronal populations could be modeled in the following way. The two lower layers (infra-granular and granular) could be an exact copy of the two-layer model presented here. The supra-granular layer would be yet another copy of the upper layer in this work, stacked on top of the other two layers. Blood draining effects would exist from infragranular to granular layers, and from granular to supra-granular layers (the plausibility of alternative implementations could be evaluated by model comparison).

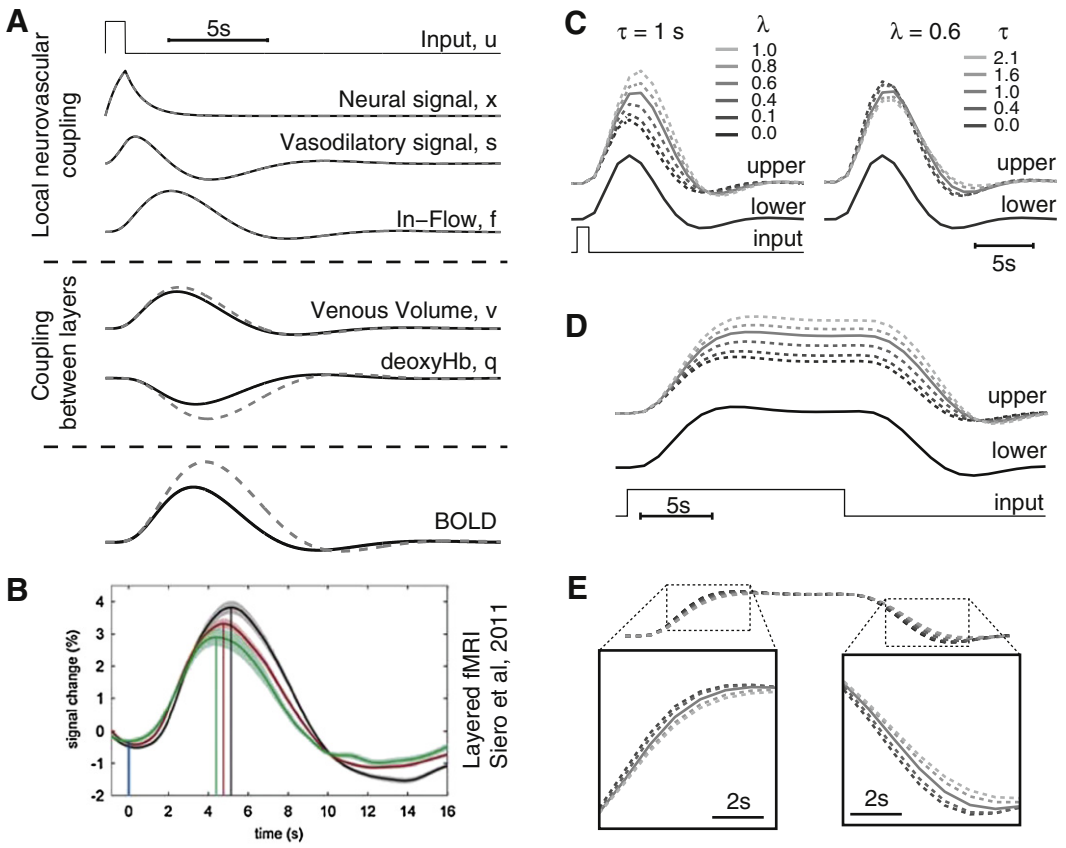

Fig. 2. Illustration of model dynamics and effect of parameters. A) Response of all variables to a brief input of one second duration. Traces are overlaid for lower layer (solid black) and upper layer (dashed gray). Note that the blood draining only affects v and q, as well as the resulting BOLD response. B) Hemodynamic response function for three cortical depths (3–2 mm: green; 2–1 mm: red; 1–0 mm: black) in motor cortex. Image taken from Siero et al. (2011), reproduced with permission. Parameters of the simulation in A) were not fitted to match these experimental traces. Note the qualitatively good agreement: increased peak amplitude and delay of responses in upper layers. C) Effect of parameters λ_d and τ_d while the other parameter was kept fixed (see insets). The gray solid line in the upper layer indicates the same parameter setting as in A). D) Effect of parameter λ_d for a longer input pulse of 15 seconds duration. The gray solid line in the upper layer indicates the same parameter setting as in A). Legend for gray scale as in C (left). E) Effect of parameter τ_d on BOLD signal in the upper layer for a longer input pulse of 15 seconds duration. τ_d mainly affects the rising and falling slope of the response to stimulation (see insets). The gray solid line in the upper layer indicates the same parameter setting as in A). Legend for gray scale as in C (right).

of explained variance.[2] Furthermore, it is one of the definitions of signal to noise in fMRI considered by the comprehensive review of Welvaert and Rosseel (2013). However, SNR as defined here clearly differs from the definition often used to characterize the acquisition quality of MR images. There, the mean signal intensity is compared to the fluctuations around it. In the simulations, the noise ranged from very high SNR of 10 to low SNR of 0.5. For comparison, these values correspond to values of explained variance between 99% (SNR = 10) and 20% (SNR = 0.5). As a control, we also ran simulations using noise only (SNR = 1/1000: explained variance < 0.1%).

Prior distributions of model parameters

The prior distributions of parameters constitute an integral part of any model. In the following, we describe all priors of our models. The priors for the parameters of the standard hemodynamic response model as well as for the neuronal connections followed SPM8 (r4667; www.fil.ion.ucl.ac.uk/spm/). The two hemodynamic parameters ($\kappa_{u,l}$, $\tau_{u,l}$) were fitted separately for upper and lower layers, while only one ε was estimated during inversion. Three other parameters were equal in both layers and fixed in all simulations (i.e., prior variance set to 0): $\gamma = 0.41$, $\alpha = 0.32$ and $E_0 = 0.34$. The two newly introduced parameters describing the hemodynamic coupling across layers – time delay τ_d and coupling strength λ_d – are both confined to positive values and were thus taken as log-normally distributed. Table 1 provides a summary of all prior parameter distributions which were used in our analyses of simulated data below (unless indicated otherwise). In addition, we also provide the values ($para_{gen}$) that were used to generate the data. For analyses where τ_d was not inferred from generated data, we

Table 1
Prior distributions for hemodynamic parameters.

Parameter	μ_0	σ_0	$\mu_{phys} = exp(\mu_0)$	$para_{gen}$
$\kappa_{u,l}$	0.65	0.040	1.92	1.92
$\tau_{u,l}$	0.98	0.049	2.66	2.66
ε	−0.78	0.244	0.46	0.46
τ_d*	0	1.41	1	1
λ_d*	−2	10	0.14	0.5

Summary of prior distributions of hemodynamic parameters. The mean μ_0 and standard deviation σ_0 of the prior distributions are given in log space (their native scale during model inversion). μ_{phys} denotes the true (exponentiated) physical value which enters the likelihood function. $para_{gen}$ are the parameter values used for generating data, i.e. simulations.

* The value of these priors varied across simulations. Values differing from these priors are provided in the results section, where applicable.

[2] For a fixed HRF function with amplitude 1 and standard deviation σ_{HRF} and noise σ_{noise} the contrast to noise ratio of a signal of amplitude a is CNR = a / σ_{noise}, while the SNR as defined in this paper will result to SNR = (a·σ_{HRF}) / σ_{noise} = σ_{HRF}·CNR. The percentage of explained variance var(signal) / var(signal + noise) can be written as $\frac{\sigma^2_{signal}}{\sigma^2_{signal}+\sigma^2_{noise}} = \frac{SNR^2}{(SNR^2+1)}$, and can thus also be directly related to the SNR measure used here.

assumed a value of $\tau_d = 1$ s. This is in line with recent measurements of blood flow velocity in cortical ascending venules (1 mm per several hundred ms) (Santisakultarm et al., 2012) and assumes a cortical thickness of a few millimeters. Please note that in analyses of empirical data, due to regional differences in cortical thickness across the brain, this parameter is likely to differ among cortical areas.

Prior distributions for the neuronal parameters were set according to standards of DCM in SPM8 (r4667). Neuronal self-connections (within layers) had a negative prior mean, in order to ensure stability of the network. Priors for connections between layers and for input weights were shrinkage priors, centered on μ_0, but with a relatively large variance. Further details about priors are given in the Supplementary Tables S1 to S3.

Application to empirical data from a visual experiment

In order to test the layered hemodynamic model empirically, we applied it to previously published measured layer-wise fMRI data (Koopmans et al., 2010b). In brief, fMRI data was measured in visual cortex using a blocked design with periods of luminance-matched flickering concentric ring patterns: black-white, red-green, yellow-violet and rest (gray screen). The duration of blocks was 10 volumes. MRI data were acquired on a 3 T whole body scanner (TIM-Trio, Siemens Healthcare, Erlangen, Germany) using a 32-channel head coil. The parameters for the 3D-EPI sequence were as follows: voxel size $0.75 \times 0.75 \times 0.75$ mm^3, matrix 192×256, 32 slices covering the calcarine sulcus, TE 30 ms, TR 79 ms, flip angle 20°, bandwidth 840 Hz/pixel, 6/8 partial Fourier, and acceleration factor 4 in the primary phase encoding direction. 640 volumes were acquired with a volume TR of 2.5 s. We used in-house software (Koopmans et al., 2011) and Freesurfer (Dale et al., 1999; Fischl et al., 1999) to define two compartments – roughly corresponding to infra- and supra-granular layers – within primary visual cortex and then extracted the average time course from each of the two layers. More specifically, Freesurfer was used to find the white-matter and pial surface meshes in anatomical images which were warped to the EPI images and corrected for distortions. V1 patches were drawn on these surface meshes and for each node within V1 the line between the corresponding nodes on the white-matter and pial surface was sampled from the EPI data, creating a through cortex profile for each node. All profiles within V1 were then averaged to yield a profile time-course which was split into an upper and lower half. The average time courses of the upper and lower part represented the data to which our layered hemodynamic model was applied (for more details, see Koopmans et al. (2011)).

In this empirical analysis, we tried to find the most plausible explanation for between-layered differences in signal, considering three main possible reasons: (i) different strengths of neuronal inputs to each region, (ii) different local hemodynamics, or (iii) blood draining effects across layers. Our focus was thus on inferring the hemodynamic parameters and direct inputs; by contrast, we fixed the neuronal connectivity parameters (A matrix parameters). The within-layer self-connections were set to be highly negative ($A_{uu} = A_{ll} = -5$), leading to fast neuronal transients, and the neuronal connections between layers were set to zero. This resulted in a model where the neural activation in the two layers was completely determined by their respective input, rendering this analysis equivalent to the approach in Friston (2002). The parameters estimated in this version of the model are the input strengths to the two layers as well as the hemodynamic parameters. For all models, we represented the three stimulation conditions (contrasts of the flickering checkerboards) by separate inputs.

We varied model structure along the following dimensions: (i) no blood draining effect (λ_d fixed to 0) vs. estimation of this draining effect (λ_d as free parameter), (ii) identical input to both layers vs. independent inputs to the two layers, (iii) identical local hemodynamic parameters in both layers vs. independent hemodynamic parameters for the two layers, and (iv), fixed blood flow delay ($\tau_d = 1$ s) vs. τ_d as free parameter. Please note that estimation of τ_d is only possible for models that include blood draining. This resulted in a model space comprising 12 different models (Table 2). For later family model comparison the 12 models were subdivided into 4 families according to whether blood draining was included in the model and whether local hemodynamic coupling was forced to be identical or was allowed to differ across the two layers (cf. Table 2). Families 1 and 2 did not include blood draining while families 3 and 4 had it included. Families 1 and 3 had the same local hemodynamic parameters for both layers, while in families 2 and 4 the local hemodynamics were allowed to differ between layers.

Table 2
Model space for application to experimental data.

Model Nr.	1	2	3	4	5	6	7	8	9	10	11	12
Family Nr.	1	1	2	2	3	3	3	3	4	4	4	4
BD					x	x	x	x	x	x	x	x
#hp	1	1	2	2	1	1	1	1	2	2	2	2
τ_d							x	x			x	x
#Inp	2	1	2	1	2	1	2	1	2	1	2	1

BD: blood draining (included); #hp: uniform (1) or layer-specific (2) hemodynamic parameter sets; τ_d: delay constant estimated; #Inp: number of inputs.

At the present time, we have little empirical knowledge about inter-regional and inter-individual variability in the hemodynamic parameters considered here, and the current choice of prior variances may not be optimal. A wider prior conveys more flexibility in fitting a particular parameter while endowing the model with higher complexity. Thus, inference on most likely mechanisms through model comparison (e.g., differences in local hemodynamics across layers vs. blood draining) is influenced by the relative width of priors for the respective parameters. Here, we included the relative width of the priors as an additional factor in model space. We varied the prior variance of the two local hemodynamic parameters in four steps from small to large (cf. Table 3) allowing for increasing flexibility of the parameters of local hemodynamics. The entire model space thus contained 48 models in total.

We used the standard variational Bayesian approach for model inversion in SPM (Variational Laplace), with adapted priors as described above, to fit the 48 models to the empirical data of 10 subjects. Bayesian model selection (BMS; Penny et al., 2004; Stephan et al., 2009), based on a free-energy approximation to the log model evidence, was used to evaluate the relative goodness of competing models. Critically, the log-evidence does not simply reflect the fit of each model, but its trade-off between accuracy and complexity, thus shielding against overfitting. Furthermore, family-level model comparison (Penny et al., 2010) can be used to compare sets of models which differ along a particular structural dimension, enabling one to integrate out uncertainty about detailed aspects of model definition and assessing the relative importance of general mechanistic dimensions. Here, family-level BMS was employed to conduct two critical comparisons. First, we compared models including blood draining vs. models without blood draining, in order to assess whether taking into account blood draining effects was

Table 3
Priors for hemodynamic parameters: application to data.

Parameter	μ_0	σ_0	$\mu_{phys} = exp(\mu_0)$
$\kappa_{u,l}$ [+]	0.65	0.04, 0.30, 0.81, 2.19	1.92
$\tau_{u,l}$ [+]	0.98	0.05, 0.36, 0.99, 2.69	2.66
ε (3 T)	−0.78	0.24	0.46
τ_d [*]	0	1.41	1
λ_d	−0.69	1.41	0.5

Other hemodynamic parameters were kept at fixed values: $E_0 = 0.34$, $\gamma = 0.41$ and $\alpha = 0.32$.

[+] All four different standard deviations ($\sigma_1, \sigma_2, \sigma_3, \sigma_4$) spanning the model space are provided.

[*] τ_d was kept fixed, i.e. $\sigma_0^2 = 0$ for models where τ_d was not inferred.

important at all for explaining the measured data. Second, we compared four families of hemodynamic models (one vs. two sets of hemodynamic parameters, as well as models with vs. without blood draining). This second comparison allowed us to test, in addition to blood draining effects, whether models with identical or different local hemodynamics for the two layers provided a better explanation of the data.

Results

In the following, we first challenge the model in a series of simulations. The simulations illustrate how taking into account blood draining effects leads to notable changes in the BOLD signal of the upper layer. Model comparison based on simulated data is then used to test to what extent it is possible to differentiate neuronal connections between layers from blood draining. In addition, we show simulations that illustrate to what degree inputs to different layers can be disentangled. Further simulations investigate how well known parameter values can be estimated using model inversion. Finally, we focus on the hemodynamics only and demonstrate that in a simple visual paradigm blood draining explains the observed data better than layered differences in local hemodynamics.

Qualitative properties of simulated layer-wise BOLD data

The effect of the blood draining from lower to upper layers is illustrated by showing simulated BOLD responses of a two-layer model for different settings of two parameters λ_d (draining strength) and τ_d (draining delay). Both layers received exactly the same input and had the same intrinsic connectivity and local hemodynamic parameters. Thus, in the absence of any coupling, the BOLD response in both layers would be identical. Fig. 2A shows the time course of all variables within the two-layered model in response to a brief input of 1 second duration. The parameters for this simulation were set to $\lambda_d = 0.6$ and $\tau_d = 1$ s. Because the draining effect concerns only volume and deoxyHB changes, the neuronal signal (x), the vasodilatory signal (s) and the in-flow (f) are identical in both layers. The modeled BOLD responses are qualitatively in excellent agreement with experimentally measured responses (Fig. 2B), for example in Siero et al. (2011). The effect of the parameters λ_d and τ_d is illustrated in Fig. 2C. Finally, Fig. 2 (D and E) shows the response to a prolonged input of 15 seconds duration. These simulations illustrate that λ_d mainly affects the signal amplitude in the upper layer, while τ_d changes the behavior at the transients of the input.

Model comparison — distinguishing effects of neuronal connectivity and blood draining

One of the main purposes of generative models like DCM is to compare alternative models representing competing hypotheses of how the data were caused. Here, we applied model comparison to simulated fMRI data in order to test whether across-layer influences caused by blood draining and neuronal connectivity, respectively, can be disambiguated by our model. Specifically, we simulated data using the two models illustrated in Fig. 3A where the two layers received different inputs as described in the Methods section. The blood draining (BD) model did not have any neuronal connection between the two layers ($A_{ul} = 0$), but the lower layer influenced the upper layer via blood draining ($\lambda_d = 0.61$). The neuronal connectivity (NC) model did not have any blood draining ($\lambda_d = 0$) but an excitatory connection from the lower to the upper layer ($A_{ul} = 0.5$). This scenario was chosen in order to make the distinction between BD and NC models as difficult as possible. A full microcircuit with reciprocal connections between upper and lower layers will be considered below.

Fig. 3B shows sample BOLD traces for the two models at a noise level of SNR = 3. Note that the parameter values for generating data were chosen such that the simulated BOLD traces were highly similar for

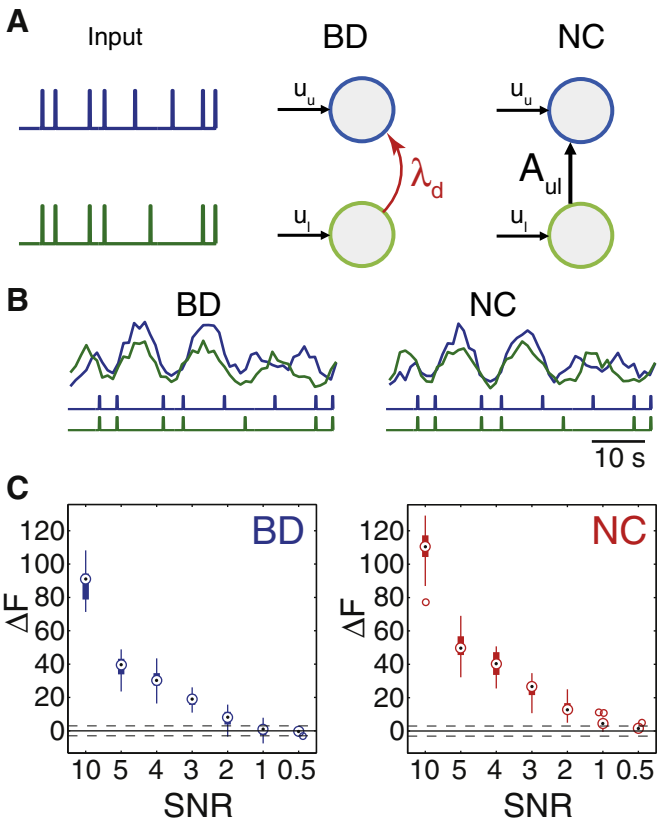

Fig. 3. Comparison of forward neuronal connection and blood draining. A) Illustration of the two models used for generating data. Inhibitory neuronal self connections were included in the models, but are not displayed. B) Simulated traces for the two coupling conditions: blood draining (BD) and neuronal connectivity (NC). Lower layer (green) and upper layer (blue) traces are overlaid. SNR = 3. C) Free energy differences (ΔF) for comparing the true (BD or NC) against the alternative model (NC or BD) as a function of signal-to-noise ratio (SNR). Positive values larger than 3 (threshold indicated by broken lines) indicate that the correct model can be identified with strong confidence (Bayes factor > 20). Zero indicates lack of model discriminability. Box plots illustrate the median (black dot) and the 25th to 75th percentile range (box). Whiskers show the total range of the data with single outliers shown as circles. The model comparison was performed for 25 simulated datasets differing only in the randomization of the noise.

the two models. The similarity of the responses reflects the difficulty of this model comparison challenge. We then applied each model to each synthetic dataset, using the priors given in Table 1 and Supplementary Table S1 (we did not infer the time constant τ_d in these analyses but allowed local hemodynamics described by $\kappa_{u,l}$ and $\tau_{u,l}$ to differ between layers), and computed the negative free energy as an approximation to the log model evidence. To indicate how well the two models can be separated on average, given a single measurement (simulation), we plot the distributions of (negative) free energy differences (ΔF) for different SNR levels. Both models could be distinguished very clearly (with $\Delta F = \log BF > 3$) up to a noise level of SNR = 2 (Fig. 3C). The NC model was correctly identified even for most simulations with SNR = 1. Importantly, the free energy difference approached zero for higher noise levels. Hence, model comparison did not result in selecting the wrong model; instead, the two models could no longer be clearly distinguished for SNR < 2. For classical analysis, it has been suggested that onset times of the BOLD response can be used to distinguish local from drained effects (Siero et al., 2011; Siero et al., 2013). In particular, a response with an onset time shorter than the expected draining time should be of local origin.

In many cases, a more realistic – but even more challenging – scenario is a fully connected model with reciprocal connections between upper and lower layers, allowing for reverberating activity across laminae. We simulated such a more complex network and again tried to

separate the effects of hemodynamic and neuronal coupling by model comparison. In particular, we compared a model that had increased neuronal connectivity from the lower to the upper layer (NC) with a model that included blood draining between the two layers (BD), while assuming reciprocal neuronal connections between layers in both cases. As above, parameter values for generating data were chosen such that the simulated BOLD traces were highly similar for the two models (see Fig. 4B). Supplementary Table S2 summarizes the main parameters of interest for this simulation. All other priors were set according to Table 1 and Supplementary Table S1.

Even though the neuronal dynamics is much richer in this setting with reciprocal neuronal connections, the two different across-layer mechanisms could still be separated by model comparison, albeit at a slightly higher SNR than in the simpler feedforward case above. Please note that in either model, the neuronal connection A_{lu} from the upper to the lower layer was a free parameter. The detailed results of this model comparison are shown in Fig. 4.

Model comparison — distinguishing different sources of layered input

When investigating layered responses, distinguishing inputs to upper and lower layers may be of particular interest in future applications. This is because layer-specific connections play an important role in computational concepts such as "predictive coding" in which connections from infragranular and supragranular layers signal predictions and prediction errors, respectively (Bastos et al., 2012). We thus generated data using a model that included blood draining and subsequently tested whether accounting for blood draining effects in models applied to the synthetic data had an effect on how successfully the correct distribution of inputs was recovered. Two event related input trains (u_u and u_l) differing for upper (u_u) and lower (u_l) layer were used to generate the data (see Methods). We considered two scenarios: One without connections between the layers (Fig. 5A) and one with recurrent connectivity (Fig. 5B). Supplementary Table S3 summarizes the parameters for the generating models as well as the priors used for inversion. Here, we assumed that local hemodynamic coupling was the same in the two layers. We simulated 30 instances of each model with different instantiations of noise. In order to examine how well the two mechanisms could be disambiguated in the typical experimental setting of a group study, random effects model comparison was then used to compare two models, the generating model vs. an alternative model with the lower layer input also added to the upper layer (cf. Fig. 5). In order to demonstrate that including blood draining effects in models of layered fMRI strongly affects the inference on input

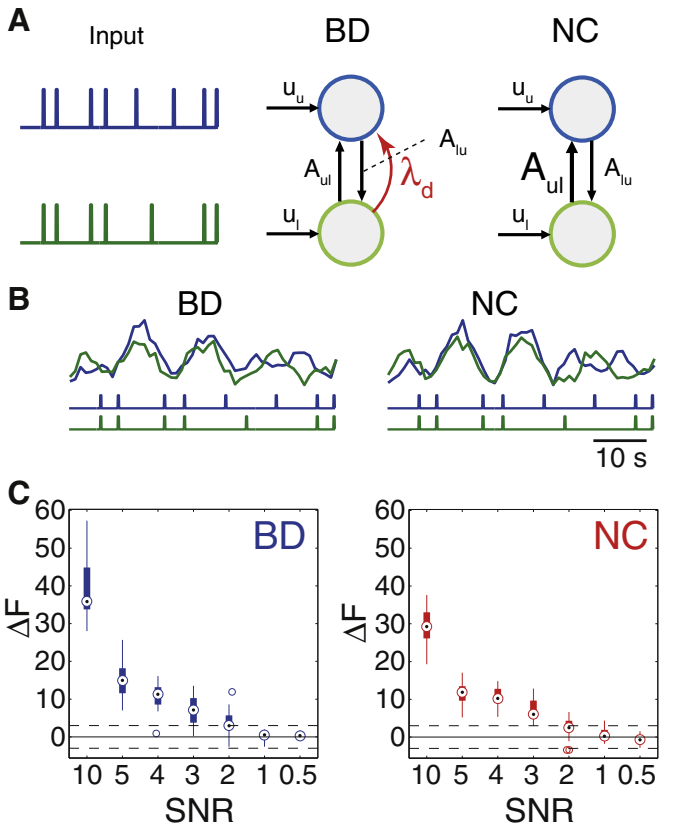

Fig. 4. Comparison of recurrent neuronal connection and blood draining. A) Illustration of the two models used for generating data. Inhibitory neuronal self connections were included in the models, but are not displayed. B) Simulated traces for the two coupling conditions: blood draining (BD) and neuronal coupling (NC). C) Free energy differences (ΔF) for comparing the true (BD or NC) against the alternative model (NC or BD) as a function of signal-to-noise ratio. Positive values larger than 3 (threshold indicated by broken lines) indicate that the correct model can be identified with strong confidence (Bayes factor > 20). Zero indicates lack of model discriminability. Box plots illustrate the median (black dot) and the 25th to 75th percentile range (box). Whiskers show the total range of the data with outliers shown as circles. The model comparison was performed for 25 simulated datasets differing only in the randomization of the noise.

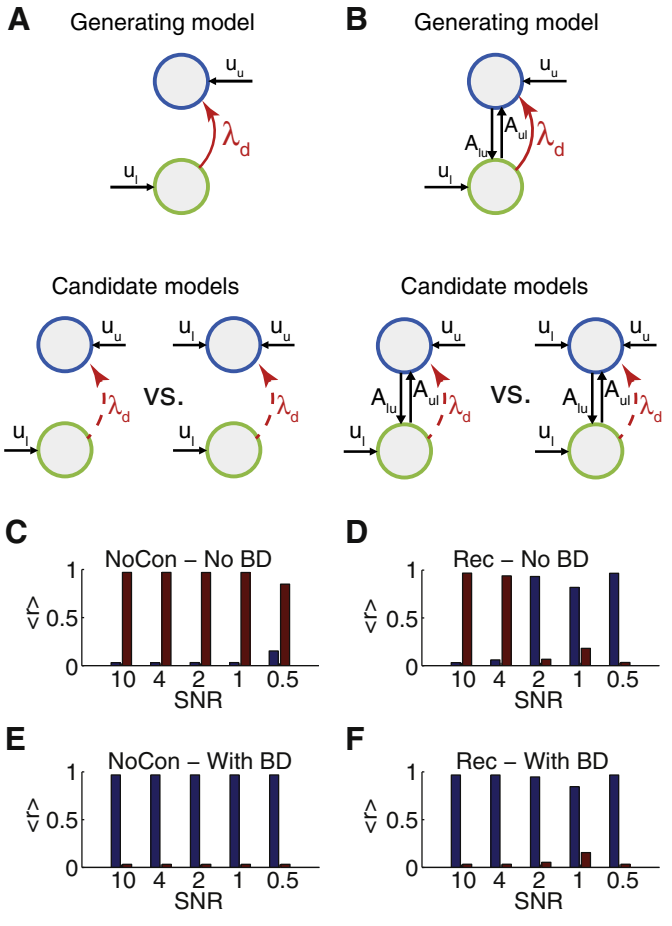

Fig. 5. Inferring the input structure to a layered circuit. A) and B) Illustration of the models used for generating data (top) and the two candidate models (bottom) used to infer input structure. Connections between layers were included only in B). Blood draining was included when generating the data, but inference was made with two separate settings (with and without) blood draining, as indicated by the dashed red arrow. Inhibitory neuronal self connections were included in the models, but are not displayed. C) and D) Model comparison results without blood draining for the models without (C) and with (D) recurrent neuronal connections. Expectation of posterior probabilities of the correct (blue) and alternative (red) model are shown for different levels of noise. E) and F) Same as in C and D but for models including the blood draining effect. All exceedance probabilities were $p_{exc} > 0.99$ in favor of the model with the higher expectation of the posterior.

distribution, we compared these two models under two different conditions, with and without blood draining. In particular, we expected the model which (wrongly) assumes lower layer input targeting also the upper layers to explain the data better, if blood draining effects were not included in the candidate models. Figs. 5C and 5D compare the expectation of the posterior probability of the correct (blue) and alternative (red) model without blood draining. Note that the data was generated with blood draining. Clearly, not including blood draining dramatically changes the inference, even when neural connections (as in the recurrently connected model, 5D) could, in principle, account for an effect of the lower on the upper layer. Consequently, the model with recurrent connections correctly infers the inputs for higher noise scenarios, where the distinction between blood draining and neural connections is difficult (cf. Fig. 4). When blood draining was included, as in the generation of the data, the correct model was always chosen (Fig. 5E and F).

Parameter estimation — inferring the parameters of the generating model

An additional question we addressed is how accurately known parameter values can be inferred from noisy data. To test this, we investigated two different scenarios. First, each of the three parameters A_{ul} (neuronal coupling strength), λ_d (blood draining strength) and τ_d (blood draining delay) were inferred separately (with the other two parameters fixed). Second, we investigated how well the two coupling parameters A_{ul} and λ_d could be inferred simultaneously for a fixed τ_d of 1 s. Note that the parameters for the local hemodynamics ($\kappa_{u,l}$, $\tau_{u,l}$) were inverted simultaneously in all these simulations, further increasing difficulty. All simulations for parameter estimation are illustrated at a noise level of SNR = 3.

Individually, the three parameters describing the blood draining (A_{ul}, λ_d, and τ_d) were estimated fairly robustly from the simulated data (see Fig. 6) for a wide range of parameter settings. In particular, inferring the time constant τ_d is not easy, as it mostly influences the on- and off-transients of the hemodynamic response, but not the overall activation in the upper layer (compare Fig. 2E). Please note that the nature of Bayesian inference has the general consequence that parameter estimates are a weighted compromise between the prior and the data (likelihood), where the weighting depends on the relative precisions. That is, in simulations where the "true" (generating) values of parameters differ non-trivially (relative to the prior precision) from the prior mean, it is expected that the ensuing posterior parameter estimates deviate from the parameter values used to generate the data. This effect reflects the regularizing influence of the prior and is also referred to as "shrinkage"; it can be observed e.g. in Fig. 6C. This effect grows with increasing distance of the generating parameter value from the prior mean (and additionally depends on the form of the likelihood function, e.g., the degree of conditional dependencies among parameters that it induces).

Next, we tested to what degree the two parameters that control the strength of the influence of the lower layer on the upper layer – neuronal coupling and blood draining parameters – could be inferred simultaneously. Importantly, these are the parameters which we anticipate will be most relevant for future applications of our model to empirical fMRI data, i.e., analyses of layer-specific inter-regional connectivity. By contrast, τ_d is not a parameter of major interest for analyses of connectivity because, in contrast to the other parameters, empirical measurements and model based estimates exist (Boas et al., 2008; Santisakultarm et al., 2012). This allows for treating this parameter as relatively well known and either fixing it (e.g., $\tau_d = 1$ s in the following simulations) or using a very tight prior. This reduces the problem of inference to neuronal coupling and blood draining parameters, in order to explain differences in layer-wise hemodynamic responses that may otherwise confound estimates of inter-regional connectivity estimates.

Here, we thus generated data with a two layer model that had both blood draining as well as an excitatory neuronal connection. Parameters were changed on a grid using values A_{ul} = [0 0.25 0.5 0.75 1] and λ_d = [5e-5 0.14 0.37 0.61 0.78 1] and all priors (except for τ_d) set according to Table 1 and Supplementary Table S1. The simulations show that the two

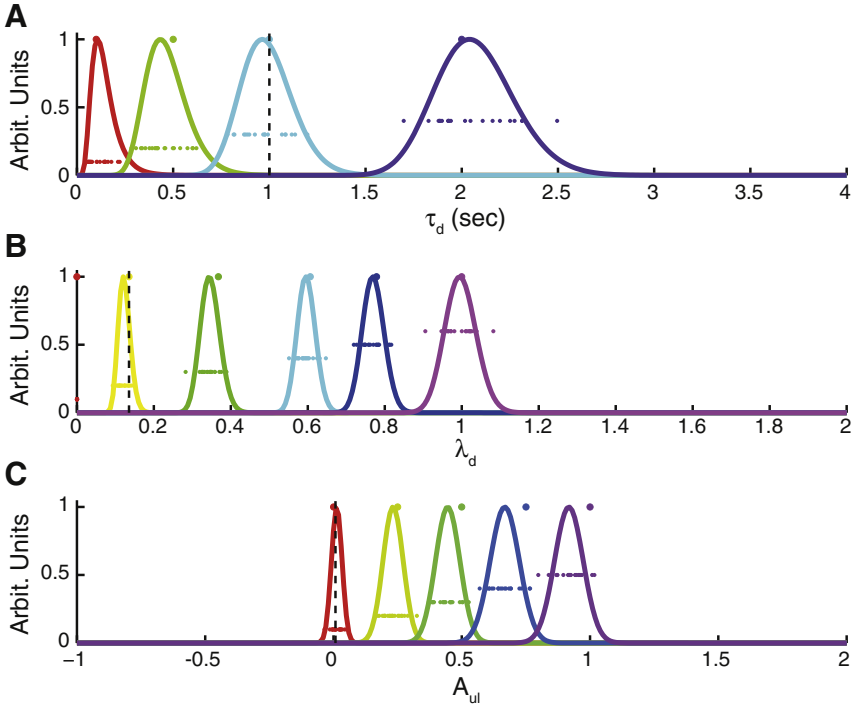

Fig. 6. Posterior parameter estimates when inferring the three key parameters of the across-layer coupling separately: A) blood draining delay τ_d, B) strength of the blood draining effect λ_d, C) excitatory connection from lower to upper layer A_{ul}. Colors indicate results for different generating parameter values (indicated by the dots on top of the curves). The curves represent the sample distributions of the maximum a posteriori estimates over 20 simulations (colored dots) with independent noise, for the different parameter values (all simulations SNR = 3). Dots are plotted at different heights so that neighboring distributions can be disentangled more easily. Vertical broken lines indicate the mean of the prior distribution which was kept identical for all model inversions.

parameters could be inferred robustly in most cases; Fig. 7 provides a visualization of these results and highlights cases were parameter values were underestimated, probably due to a joint influence of the prior and the strong negative conditional dependency between the two parameters. The latter is expected, given that both parameters contribute to how strongly the lower layer influences the upper layer. However, the blood draining occurs on a time scale that is slow enough to be distinguished from the neuronal connection.

Application to real data

Finally, we set out to test the model of layered hemodynamics in an application to real data. The fMRI data used for this application were recorded while 10 subjects viewed blocks of flickering checkerboards and have been presented previously (Koopmans et al., 2010b). In this analysis we focused on the hemodynamics only and thus fixed all neuronal connection parameters: There were no connections between layers and a fixed, fast decay within layers. Thus, only activation amplitude, defined by the input weight and hemodynamic parameters were estimated, similar to the analyses in Friston (2002). We compared 12 different models (cf. Table 2) with different versions of local hemodynamics and hemodynamic coupling. In addition, we varied the flexibility of the hemodynamic parameters ($\kappa_{u,l}$ and $\tau_{u,l}$) by including 4 different levels of variance for the respective priors. We then inverted all 48 resulting models and used Bayesian model family comparison (Penny et al., 2004; Stephan et al., 2009; Penny et al., 2010) to assess the relative goodness of the models. Specifically, in order to investigate the importance of blood draining, we compared families of models with and without BD. Families with BD clearly outperformed models without BD (expected probability of the model $p = 0.92$, exceedance probability $p_{exc} > 0.99$, Fig. 8B). Next, we compared four families, defined by different combinations of present vs. absent blood draining effects and uniform vs. layer-specific hemodynamic parameters (cf. Table 2). Again, the family comparison showed a clear winning model family (Fig. 8C); this family 3 contained models with uniform hemodynamic parameters across layers and did account for blood draining effects ($p_{exc} = 0.96$). The posterior probabilities for the 4 families were 0.07 (family 1), 0.07 (family 2), 0.68 (family 3) and 0.18 (family 4). Finally, when examining the 12 models individually (Fig. 8D, family comparison over the 4 settings for hemodynamic priors), models number 8 (expected $p = 0.29$, $p_{exc} = 0.65$) and 6 (expected $p = 0.16$, $p_{exc} = 0.24$) outperformed the other models (all expected $p < 0.09$, all $p_{exc} < 0.04$). Both these models included one set of hemodynamic parameters for both layers and accounted for blood draining. The best models had intermediate variance (3rd level, cf. Table 3) for the hemodynamic priors. τ_d was a free parameter in model 8 and was estimated to be on the order of several hundreds of milliseconds (mean ± std: 641 ± 136 ms). This estimate is highly similar to prior experimental results, where the delay in peak time was roughly 0.22–0.24 s per mm when moving from deep to superficial layers (Siero et al., 2011). Insets in Fig. 8D illustrate the prior distributions and maximum a posteriori estimates of the parameters for the two winning models at σ_3 from all 10 subjects.

On average, the model was able to capture the fMRI traces well in both layers. Fig. 8E shows example traces (averaged over 16 cycles) of two subjects. The same plots are given for all subjects in Supplementary Fig. S1. The model accounted for a considerable amount of the variance of the data, as demonstrated by SNR values between 1.27 (61.7% variance explained) and 0.53 (21.9%) in the upper (median 0.92 (45.8%)) and 1.11 (55.2%) and 0.43 (15.6%) in the lower layer (median 0.80 (39.0%)). These SNR values were calculated over the entire time series, which was also used for modeling. It should be noted in this context that our variational Bayesian optimization procedure for fitting the models does not strive for maximizing fit, but instead optimizes the balance between model fit and model complexity (where the latter includes, among other things, the deviation of posterior parameter estimates from the prior mean). This prevents overfitting and maximizes generalizability (for details, see Stephan et al., 2009).

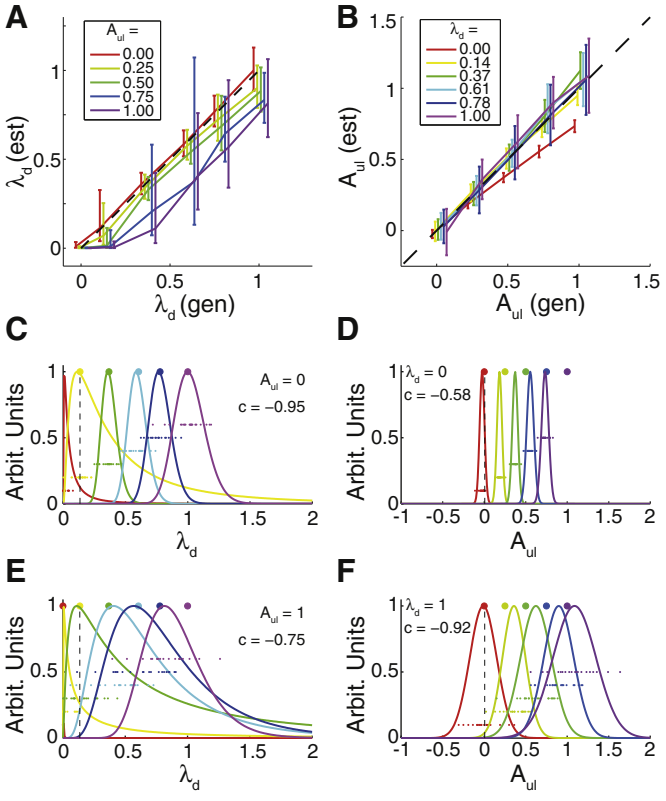

Fig. 7. Estimating blood draining strength and neuronal connection strength simultaneously. A) and B) show the posterior mean of the parameter estimates for λ_d and A_{ul}, respectively. Both are plotted as a function of the true generating parameters. Colored lines show all simulations for a particular strength of the other parameter (see insets). Error bars correspond to one std in the native space of the parameters and are thus non-symmetric for λ_d (which is estimated in log-space). The black dashed line indicates the true parameter values. C) and E) show the posterior parameter estimates for λ_d of all simulations for the case of no neuronal connection (C, $A_{ul} = 0$) and strong neuronal connection (E, $A_{ul} = 1$). See insets. Colors indicate different values of the true value of the parameter, indicated by the dots on top of the curves. Compare also inset in Panel B. D) and F) show the posterior parameter estimates of A_{ul} for all simulations for the case of no blood draining (D, $\lambda_d = 0$) and strong blood draining (F, $\lambda_d = 1$). Colors indicate different values of the true value of the parameter, indicated by the dots on top. Compare also inset in Panel A. The average posterior correlation c between λ_d and A_{ul} (mean over all simulations in the graph) is indicated by the insets in C–F.

Discussion

High-field fMRI at resolutions below 1 mm poses analysis problems that can differ fundamentally from standard fMRI analysis. Here, in the context of laminar fMRI, we have presented an extended version of a commonly used hemodynamic response model (Buxton et al., 1998; Friston et al., 2000; Stephan et al., 2007). The model phenomenologically captures effects of venous blood draining from lower to upper layers of cortex. The simulations reproduce observed layered BOLD responses in visual and motor cortex in a qualitative manner (Siero et al., 2011). Model inversion (based on simulated data) showed that the effects of blood draining and neural connectivity can be separated, that including blood draining strongly affects inference on the layered input structure to a microcircuit, and that it is possible to infer the generating parameters. An application to a visual fMRI data set revealed that the proposed model captures the measured responses better than the compared models with more flexibility on the local hemodynamic coupling parameters. Below we discuss the results and outline the consequences for future research of layered cortical circuits based on high-resolution fMRI.

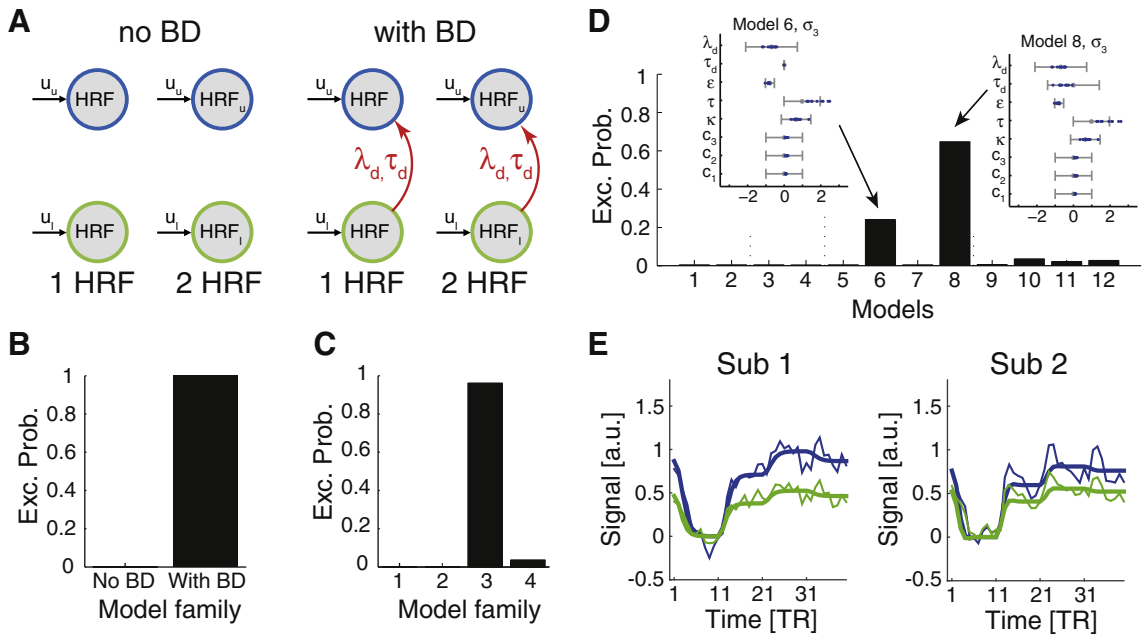

Fig. 8. Model family comparison for visual fMRI data. A) Illustration of the four model families with the two factors 'no BD vs. BD' (no blood draining included vs. blood draining included) and '1 HRF vs. 2 HRF' (one single set of local hemodynamic parameters vs. two independent sets of local hemodynamic parameters for the two layers). Models within those families differed with regard to whether the input strengths u were estimated separately for each layer or not, and whether the delay τ_d was estimated or not (for models including blood draining only). B) Family comparison between models with and without hemodynamic coupling. C) Family comparison over 4 families as defined in Table 2. Note that families 3 and 4 include BD, while families 1 and 2 do not. D) Model comparison results (exceedance probability) for all 12 models. Note that each model here consists of 4 instances with varying priors for the hemodynamic parameters (cf. Table 3). Dotted lines indicated borders between families (1–4 from left to right). Insets show the maximum a posteriori estimates for the parameters of the two models with highest model evidence. Note that with exception of the input weights (c_i) all parameter estimates are shown in log-space (and are thus negative for parameter values smaller than one). E) Measured BOLD signals from two example subjects (all remaining subjects are shown in Supplementary Fig. S1). The plots show the average (thin line) fMRI signal over one cycle of stimulation (average taken over all 16 cycles of the scanning session). Bold lines indicate the fit obtained by the winning model (cf. D). Colors indicate lower (green) and upper (blue) layer. Time axis is scaled to units of TR. Visual stimulation changed every 10 TRs at the indicated time points 1, 11, 21 and 31.

Cortical blood supply and hemodynamics

The architecture of the blood vessel system in cortex (Duvernoy et al., 1981; Weber et al., 2008) is highly complex. However, one of the hallmarks of the vascular network within cortex is the arrangement of arteries (arterioles) and veins (venules) perpendicular to the cortical surface. This arrangement will clearly influence the interpretation of layered BOLD signals due to venous blood from deeper layers passing through more superficial ones. In this work, we have presented a model of layered hemodynamics that focuses on effects of venous blood draining perpendicular to the layers back to the large surface vessels at the pial surface. The proposed model provides a mechanistic explanation of several features of layered BOLD signals observed in human high-resolution fMRI (Siero et al., 2011), but also in animal studies (Herman et al., 2013).

Previous work on layer-wise differences in BOLD signal has focused on other features of cortical vasculature, such as differences in capillary density across layers (Weber et al., 2008) which likely contribute to layer-specific BOLD signal, e.g., in layer 4 (Koopmans et al., 2010a; Koopmans et al., 2011). Several animal studies have interpreted layered differences in the BOLD signal in terms of layer-specific differences in the local hemodynamic coupling (Goense et al., 2012; Herman et al., 2013) and there is evidence that neurovascular coupling as measured by dilation of arterioles may vary across the cortical depth (Tian et al., 2010) as well. In fMRI, there is a longstanding discussion about effects of large superficial draining veins on the BOLD signal (Turner, 2002). Effects of blood draining have been modeled for large surface vessels, and specific non-uniform draining effects have been suggested to influence the specificity of fMRI activation patterns (Kriegeskorte et al., 2010). Interestingly, although proposed as a potential explanation for varying delays across layers (Siero et al., 2011), blood draining effects across cortical layers have been studied less systematically. In particular, current hemodynamic models do not take into account the potential impact of such blood draining effects.

In our model, two parameters control the dynamics of blood draining effects across layers. First, the time constant τ_d introduces a delay between the lower and upper layers. This time constant depends on the average blood flow velocity and on the distance between the layers, i.e. cortical thickness. Optical flow measurements in mice suggest a flow velocity of below 5 mm/s in venules up to a diameter of 60 μm (Santisakultarm et al., 2012). Similar values for the venous blood flow velocity have been reported in a detailed model of cortical vasculature (Boas et al., 2008). Our estimations based on model inversion from visual data suggest a delay of above 0.5 s, which is in the expected range given a cortical thickness of few millimeters (Fischl and Dale, 2000) and is in agreement with a measured delay of roughly 0.7 s (range 0.4 s to 1.0 s) for BOLD peak times in upper layers compared to lower layers (Siero et al., 2011). The second parameter λ_d controls how strongly the lower layer affects the upper layer. It is much more difficult to directly compare this parameter to measurements from animal experiments, and we are not aware of any study directly measuring this influence.

Comparison of modeling results with existing literature

It was not the goal of this study to provide a layered model that captures all neuronal and hemodynamic details of layer-wise activity known to date, but rather to extend the hemodynamic model for DCM of fMRI in order to specifically capture blood draining effects on layered BOLD responses. The ability of the proposed model to capture blood draining effects on layered BOLD responses has therefore been compared against the most directly related empirical findings, i.e., a detailed report of empirically measured layer-wise BOLD signals in visual and motor cortex (Siero et al., 2011). However, it may be informative to

consider this comparison in the light of the wider literature on layered hemodynamics.

Based on measured timing differences in the BOLD signal, the onset time has been suggested as a potential measure to disentangle local hemodynamic effects from blood draining (Siero et al., 2011; Siero et al., 2013). Onset times were used to uncover layered inputs using line scanning fMRI in a mouse model (Yu et al., 2014). Here, we have not studied onset times systematically. However, we also see a temporal delay of the BOLD signal in the model in upper layers compared to lower layers. For example, in Fig. 2A the peak time in the upper layer is delayed by approx. 0.5 s. This is within the range of many previous studies (Jin and Kim, 2008; Tian et al., 2010; Siero et al., 2011). But see Hirano et al. (2011) for the opposite finding in rats. Furthermore, the amplitude of the BOLD signal predicted by our model increases towards the pial surface. Previous studies in humans (Koopmans et al., 2010a) as well as in monkeys (Chen et al., 2013) are in line with this finding but have measured an additional peak of the layered BOLD amplitude in the middle of visual cortex (presumably around layer 4). The two layered model presented here is not able to capture this effect which is probably due to increased neuronal and/or local hemodynamic activity around layer 4.

In our model, CBV changes are highest in upper layers. The empirical findings on the magnitude of cerebral blood volume changes across layers are mixed. Animal studies also suggest an increase of relative venous CBV changes towards the pial surface (Kim and Kim, 2011). By contrast, arterial CBV changes, which are not represented by our model, are highest around the middle of cortex, i.e. layer 4 (Kim and Kim, 2011), resulting in relative changes of total CBV that are highest around the middle of the cortical sheet (Jin and Kim, 2008). However, see Hirano et al. (2011) and Herman et al. (2013), who report peak total CBV responses increase towards the pial surface, as in our model. Measurements in humans, finally, suggest no difference between layered CBV changes for cortical excitation (Huber et al., 2014; Huber et al., 2015).

There is evidence for a decoupling of BOLD signal and CBV in monkeys that differs between layers, in particular for inhibition (Goense et al., 2012). Similarly, the time course of the BOLD signal and CBV is decoupled in humans (Huber et al., 2015). The volume changes are delayed with respect to the BOLD response, suggesting that cerebral blood flow and CBV are decoupled. Such a decoupling cannot be achieved with the standard hemodynamic model in DCM, but an adaption of DCM for fMRI that resolves this limitation has been suggested recently (Havlicek et al., 2015).

Finally, there are layered differences in the initial dip as well as the post-stimulus undershoot of the BOLD signal (Siero et al., 2015). These differences have not been addressed with the current model, and accounting for them properly might require a more flexible implementation of the Balloon model (Havlicek et al., 2015).

Limitations of our model

The modeling results suggests that it is, in principle, possible to separate hemodynamic and neuronal components of across-layer influences. While a simple difference in the amplitude of the signal in the upper layer can always be captured by an increase in neuronal connection strength, it is the additional delay of the hemodynamic effect that enables the differentiation between neuronal connectivity and blood draining effects. The effect of this delay is only visible in transients (fast changes of neuronal activity during stimulation), as demonstrated by the simulations in Fig. 2. This dependence on transients is what makes the separation of NC and BD intrinsically difficult and requires relatively high SNR values (SNR ≥ 2) in our simulations; for analyses of empirical data, this suggests the importance of optimized data acquisition, long measurement times and experimental designs that explore dynamics of transients rather than studying long blocks, only. An alternative could be to include other aspects of BOLD transients. In rats, it has for example been shown that vasodilation in arterioles is fastest in deep layers and lags behind in upper layers (Tian et al., 2010). This leads to an increased initial dip in the BOLD response. A recent study has replicated this finding in humans (Siero et al., 2015). The authors suggested that measuring the initial dip across layers could be a promising alternative to "conventional" BOLD fMRI of cortical layers.

It is straightforward to directly compare the noise levels of the simulated data (ratio of standard deviations of true signal and independent Gaussian noise) to values of percentage of explained variance and thus to F-values. A clear distinction between the BD and NC models is achieved for SNR = 2 and above; this corresponds to a percentage of explained variance of 80%. Please note that here the SNR is calculated based on the simulated data and not on the inferred model. Hence, it relies on knowing the true signal, which is impossible in real experiments. Nevertheless, similarly high values of explained variance can be obtained with fMRI under suitable conditions, e.g., in visual cortex by population receptive field modeling (Dumoulin and Wandell, 2008). It is important to note that in addition to the parameters which mediate across-layer effects, we also inferred the rate constant of the vasodilatory signal decay (κ) and transit time of the Balloon model (τ) of local (within-layer) hemodynamics, separately for the two layers. This takes into account potential layer specific differences in local hemodynamics. However, it makes the inference of the blood draining effects even more challenging as hemodynamic responses that are specific for the upper layer could potentially be partially accounted for by adapting the across-layer hemodynamic coupling.

In all simulations, we have used non-informative priors for blood draining effects, making the distinction between the two tested scenarios more difficult. The sensitivity of future versions of the model could be improved by incorporating data on simulated hemodynamics (Boas et al., 2008; Reichold et al., 2009) in order to constrain the priors. Such a more constrained model might improve the distinction between different types of neural coupling and would also help to reduce the high correlations between the parameters. Alternatively, high temporal sampling during fMRI data acquisition should improve the model's ability to distinguish neural and blood draining effects as well, since this would capture neuronal transients better.

Our model focuses on effects across cortical layers; by contrast, it does not accommodate effects due to drainage parallel to the cortical surface, e.g., in large pial veins. An important empirical question to be addressed by future studies is to what extent the size of veins influences fMRI signals in upper layers. The current model assumes the draining effect to be homogeneous across a cortical area, not taking into account that blood flow is concentrated to veins which only constitute a small part of the tissue. In addition, it assumes that draining is perpendicular to the cortical surface and does not take into account the different size of venules in lower and upper layers. The relatively large increase in deoxyHb (Fig. 2) that is necessary to account for the experimentally observed BOLD effects might, in part, be attributable to different sizes of vessels across layers; these can only be taken into account by the model by adjusting the magnitude of change in the oxygenation level of the blood. However, a model that includes the effects of vessel size or captures micro- and macro-vascular separately would become considerably more complex than the present formulation and have to rely on much more detailed information about the distribution of blood flow, such as the average diameter of blood vessels in different cortical areas, than is presently available. Commonly used gradient echo (GE) echo planar imaging (EPI) has been successfully used to measure layered signals (Koopmans et al., 2010a; Siero et al., 2011). However, the GE signal originates mainly from larger vessels and thus its size is also linked to the venous architecture of cortex. The increased GE signals observed in upper layers could be partly due to an increased average size of intracortical veins closer to the cortical surface. Indeed, it has been shown that the GE signal in low layers, but not close to the surface, and the overall SE signal have a very similar shape (Siero et al., 2013). This suggests that in lower layers the contributions to the GE signal come mainly from microvasculature (but see Yu et al., 2012), while in

upper layers macrovascular signals contribute significantly. In order to reduce the contribution of draining veins, alternatives to gradient-echo BOLD have been proposed. T2-weighted (spin-echo) methods have been shown to be less sensitive to veins than gradient-echo (Parkes et al., 2005; Yacoub et al., 2008) even though the T2* weighting of the EPI readout kernel itself reduces the effectiveness of this method for attenuating venous BOLD contributions (Goense and Logothetis, 2006). In animal studies, CBV-weighted fMRI using injection of MION particles has shown signal changes primarily at depths with high capillary density, successfully suppressing signal in pial veins (Zhao et al., 2006). A recent study has confirmed these volume-based imaging results in humans (Huber et al., 2015). While not as sensitive as BOLD, CBV-weighted fMRI could be used in combination with BOLD fMRI in order to constrain the relative volume estimates of the model.

A final limitation of our present results is that the application of our model to real fMRI data used a standard block design with the block duration being a multiple of TR. While this is efficient for estimating mean responses, it is not the ideal design to constrain estimates for the coupling delay between the two layers. Nevertheless, the resulting estimate for τ_d of roughly 0.5 s is close to values suggested by blood flow velocity in more detailed models of vascular dynamics in cortex (Boas et al., 2008).

Extensions and potential use of layered models of fMRI

We anticipate that our model will find useful application in studying the dynamics of simplified layered circuits with fMRI. Layered microcircuits are defining computational building blocks of cortex (Douglas and Martin, 2004), and models of layered microcircuits (Heinzle et al., 2007) as well as predictive coding theories (Bastos et al., 2012) assign distinct computational quantities to different cortical layers. Studying such theoretical concepts with fMRI could be facilitated with the model presented here, for example by using it in combination with computational models of trial-wise prediction errors and dynamic causal models that consider layer-specific connections. Our simulations above demonstrated that including blood draining effects will be critical for making inference on the layered structure of neuronal inputs to a cortical region. We expect such inference on layered input structures to be most robust if other parts of the model, such as the rate constant of blood draining, can be constrained by tight priors informed by experimental measurements.

The use of layered models for fMRI critically depends on several experimental and preprocessing steps which need to be performed prior to the modeling of time series. First, fMRI images need to be acquired at very high resolution since cortex is only few millimeters thick (Fischl and Dale, 2000). Second, one needs to define the layered structure of cortex prior to sampling the fMRI signal from layers or compartments. Individual cortical layers are often difficult to separate, even in histological analyses of cytoarchitectonics, and this becomes usually prohibitively difficult with fMRI. Hence, usually layers are defined by interpolation, following the segmentation of the pial surface and the boundary between gray and white matter. Relatively robust segmentation can be achieved by (indirectly) incorporating cortical curvature (Waehnert et al., 2014). Finally, partial volume effects occur during sampling and can lead to a mixture of signals from different layers. For a detailed treatment of this issue in the context of layered BOLD fMRI see Koopmans et al. (2011). Partial volume effects will reduce the differences between layers and might compromise the inversion of models that try to represent individual cortical layers. Irrespective of the structure of any neural model one might choose to model fMRI data, the challenges mentioned above always apply and should be taken into careful consideration when depth-resolved fMRI data are used for inference about cortical circuit mechanisms.

Given the above challenges and the present resolution of layered fMRI measurements, modeling a full six-layer circuit seems unrealistic at this point. The most likely applications of layered circuit models in fMRI will be simplified models focusing on few, probably two, layers. While this is clearly a limitation, a circuit that merely distinguishes infra- and supragranular compartments is the basic form of the famous "canonical microcircuit" (Douglas et al., 1989), and this structure serves for inferring hierarchies from brain wide anatomical connections in monkeys (Markov et al., 2014). Furthermore, this two-layer circuit is also used by predictive coding theories (Bastos et al., 2012).

In conclusion, we think that at the present time, a two-layer model is adequate given the constraints of layered fMRI measurements and should serve a number of interesting applications. In particular, it is sufficient to test a central postulate by predictive coding theories, i.e., that connections originating from supragranular cells should signal prediction errors to hierarchically higher areas, whereas connections with an infragranular source should signal predictions to areas lower in the hierarchy. However, there is one interesting possible extension to how neuronal dynamics are modeled. At present, the model captures the average activation very well, but misses out on some of the transient responses at the beginning of the blocks (e.g. Sub 2 in Fig. 8E). Including neuronal adaptation in a future version of the model might allow to better capture such transient overshoots.

A second possibility will be to use the refined hemodynamic model for inferring effective connectivity based on more informed prior anatomical knowledge. Cortical long-range connections but also connections to sub-cortical structures follow very specific layered connectivity patterns depending on the hierarchical relationship of cortical areas (Douglas and Martin, 2004). These principles of layered anatomical connectivity have been employed to define hierarchies of cortical processing (Fellemann and van Essen, 1991) and provide strong constraints for connectivity patterns, including directionality, in network models of layered cortical areas. This source of information is exploited, for example, in DCMs of electrophysiological data (David et al., 2006; Moran et al., 2011), and the model presented here might allow to enable this application in fMRI, too.

In our simulations, we have used model comparison to make a binary decision whether an influence from the lower to the upper layer was either due to neural connectivity or due to blood draining. In practical cases, there may be a mixture of neuronal and blood draining effects. Investigating the relative contributions of the two influences (NC and BD) requires inference on the parameters A_{ul} and λ_d; while not trivial due to correlations between the two parameters, our simulation results presented in Fig. 7 suggest that such an analysis is, in principle, possible.

In this work, we have used linear neuronal state equations in conjunction with the layered hemodynamic model to generate fMRI responses. However, the very same hemodynamic equations can be used with other models of neural activity; cf. (Friston, 2002). In the simplest case, as demonstrated by the inversion of the model on empirical fMRI data, the model can be used to estimate the average neuronal activation in cortical layers while taking into account effects of blood draining across layers. In the future, this could be refined by estimating activation across columns perpendicular to the cortical sheet, and thus getting closer to single voxel estimates of activity. The ensuing analysis would then provide activation patterns within cortical layers and could be used to study, for example, differences of top-down and bottom-up inputs in different layers using pattern classification (Smith and Muckli, 2010; Chen et al., 2011). The blood draining parameters could in principle be estimated for every column individually taking into account variability in blood draining across the cortical sheet. However, this application will require methods for a robust definition of columns and will face the challenge that the signals to be explained by the model represent averages over only a few voxels and can thus be expected to be of rather low SNR. Yet another alternative would be to use a combined EEG-fMRI approach based on more detailed circuit models that have been used for modeling of EEG (Moran et al., 2013).

An additional important target for extending the present work is the local hemodynamic model itself. Recently, Havlicek et al. (2015) have presented an extension of the classical DCM for fMRI that no longer assumes direct coupling of blood flow and blood volume, allowing to

explain experimental findings where these two signals diverge. Due to this feature and a more comprehensive biophysical interpretability, the model by Havlicek et al. (2015) represents a promising basis for future extensions of the present work. For example, combining their model with blood draining effects across layers might enable direct modeling of combined BOLD and blood volume measurements. In particular, it would be interesting to see, whether a combination of across-layer blood draining with the local hemodynamic coupling suggested by Havlicek et al. (2015) could capture several of the effects discussed in the previous section, including the diverse empirical findings on layer-wise BOLD responses, such as those related to differences in the initial dip and post-stimulus undershoot.

In this study, we have presented a phenomenological description of blood draining across layers. It relies on a simple delayed coupling of the main quantities of the hemodynamic model, relative blood volume and deoxy-hemoglobin concentration. Clearly, this coupling does not model the physics of blood flow directly, but nevertheless successfully accounts for the observed layered BOLD responses. A possible future extension would be to include a biophysical model of blood flow across layers based on a Balloon (Buxton et al., 1998) or Windkessel (Mandeville et al., 1999) model. This could eventually lead to a combination of layered local hemodynamics with a sophisticated description of inter-layer blood flow (Boas et al., 2008; Reichold et al., 2009). Such a model might be better suited to capture differences in micro- and macrovascular BOLD responses, which are observed even in deep layers (Yu et al., 2012) and will be most pronounced in upper layers, where venules tend to be larger (Duvernoy et al., 1981; Weber et al., 2008). However, given the increased complexity of such a biophysical model, it will be important to test how well it can be inverted, and what it adds over and beyond the phenomenological model presented here. In particular, if one is not primarily interested in intra-regional (neuronal or hemodynamic) effects that depend on the layered structure of cortex, but only wishes to take these effects into account for enabling better estimates of inter-regional effective connectivity, the simpler model may be sufficient. This is an issue which will have to be addressed by model comparison in future work.

Finally, the hemodynamic model used in this work is tailored to BOLD data acquired using gradient echo sequences. However, a recent study has suggested that using a 3D-GRASE sequence reduces the influence of surface veins and could increase the specificity of layered fMRI, particularly in upper layers (De Martino et al., 2013). Adapting the present model to data from spin-echo based sequences such as SE-EPI or 3D-GRASE thus represents an interesting option for future developments.

In order to validate the layered hemodynamic model, it will be of high interest and relevance to directly measure, in animals, neuronal activity across cortical layers simultaneously with ultra-high resolution layered fMRI and test whether the model is able to recover the underlying neuronal activity. In a highly specialized line scanning approach, the onset time of the BOLD signal was shown to be related to the input to different cortical layers (Yu et al., 2014). However, it remains an open question to what degree such differences can also be extracted from hemodynamic signals sampled at lower temporal resolutions.

Conclusion

Here, we have presented a novel extension of a commonly used hemodynamic model which incorporates blood draining effects across layers. This is important for robust estimation of layered neural activation. Our present model is based on a phenomenological account of blood draining across layers and may represent a first step towards future biophysiologically motivated models of layered hemodynamics. The modeling results suggest that for the distinction between neural and hemodynamics effects it is most efficient to study fast transients. Our simulations and empirical analyses indicate that this model may contribute to more refined analyses of layered microcircuits and layer-specific connections by means of high-resolution fMRI. This, in turn, would be important for probing a wide range of cortical pathologies – such as abnormalities of cortical connectivity in schizophrenia (Stephan et al., 2006) – and for testing the implications of theories of brain functions which emphasize differential roles of layer-specific connections, e.g., predictive coding (Rao and Ballard, 1999; Bastos et al., 2012).

Acknowledgements

We acknowledge support by the René and Susanne Braginsky Foundation (K.E.S.), the Clinical Research Priority Program (CRPP) "Multiple Sclerosis" of the University of Zurich (K.E.S. and S.S.R.), the Wellcome Trust [WT100092MA] (P.J.K.), the Innovational Research Incentives Scheme of the Netherlands Organisation for Scientific Research (Research Veni Grant to H.d.O.).

Appendix A. Parameter dependency on field strength

The nonlinear BOLD signal includes a set of parameters k_i which depend on magnetic field strength and on data acquisition parameters, in particular relaxation time TE (time to echo).

$$\begin{aligned} k_1 &= 4.3\vartheta_0 E_0 TE \\ k_2 &= \varepsilon r_0 E_0 TE \\ k_3 &= 1-\varepsilon. \end{aligned} \tag{7}$$

In this section, we briefly summarize the meaning of these parameters k_i and give an expected range of their values for 1.5 T, 3 T and 7 T. Many of the values used here are based on derivations by Uludag et al. (2009) but also draw on subsequent experimental data as indicated below.

First, ϑ_0 is the frequency offset at the outer surface of magnetized vessels and depends linearly on the main magnetic field strength B_0: $\vartheta_0 \cong 28.265 \cdot B_0$. Second, r_0 represents the intravascular relaxation rate as a function of oxygen saturation. The value for r_0 can be derived from the slope of the relaxation rate of blood when plotted against deoxygenation level. Such curves have been measured for human samples at 1.5 T, 3 T and 4.7 T (Silvennoinen et al., 2003; Zhao et al., 2007) and are summarized in Uludag et al. (2009). Calculating the slope of these curves at an oxygenation level of 70% and a hematocrit of 44% yields values of $r_0 \cong 15s^{-1}$ (1.5 T) and $r_0 \cong 110s^{-1}$ (3 T). The value for r_0 at 7 T could in theory be obtained by linear extrapolation from data at lower field strengths. Such an extrapolation yields a value of 325 s^{-1}, which is close to 360 s^{-1} suggested by recent in vivo measurements in humans (Ivanov et al., 2013). Based on these two estimates we suggest using a value of $r_0 \cong 340s^{-1}$ (7 T). Notably, however, the exact value of this parameter is unlikely to be important at 7 T because k_2, which is the only parameter affected by r_0, becomes nearly zero at high field strengths. This is due to a third field-dependent parameter ε,

$$\varepsilon = \frac{S_I}{S_E} = \frac{e^{-R_{2,I}^* TE}}{e^{-R_{2,E}^* TE}}$$

which represents the ratio between intravascular and extravascular MR signal. Here, $R_{2,I}^* = 1/T_{2,I}^*$ and $R_{2,E}^* = 1/T_{2,E}^*$ are the intravascular (venous) and extravascular relaxation rates, which have different field dependencies. Approximate values of these relaxation rates are available for all common field strengths (Donahue et al., 2011 and references therein). The range of observed values (cf. Table A1) can be used to determine mean and variance of prior distributions for ε, which is treated as a free parameter during model inversion. In particular, at 7 T, the value of ε is expected to be nearly zero (Uludag et al., 2009). Thus, the nonlinear term $k_2(1-\frac{q}{v})$ will become negligible in size compared to the other two terms. In Table A1, we also provide suggested values for the prior on ε. For this we assumed a uniform distribution of relaxation times over the range given in Table A1 and then numerically calculated the mean and standard deviation of the log-normal prior for ε. We assumed a TE

Table A1
Intra- and extravascular relaxation times.

B0	$T_{2,I}^*$ (ms)	$R_{2,I}^*$ (s^{-1})	$T_{2,E}^*$ (ms)	$R_{2,E}^*$ (s^{-1})	μ_ε; σ_ε
1.5 T	90–100	10–11	55–65	15–18	0.25; 0.04
3 T	15–25	40–67	35–45	22–29	−0.78; 0.24
7 T	3–7	143–333	25–30	33.3–40	−3.99; 0.83

Values for T_2^* are taken from Donahue et al. (2011). For a detailed list of references of individual experiments, see their paper. μ_ε and σ_ε are the mean and standard deviation of the suggested log-normal prior for ε. These values were calculated assuming a uniform distribution over the T_2^* values, then calculating a distribution of ε using all possible combinations of $T_{2,E}^*$ and $T_{2,I}^*$. μ_ε and σ_ε can then be directly calculated.

of 40 ms (1.5 T), 30 ms (3 T) and 25 ms (7 T), respectively. The corresponding values of ε are 1.28 (1.5 T), 0.47 (3 T) and 0.026 (7 T).

Finally, the resting oxygen extraction rate E_0 does not depend on any scanning parameters and is usually assumed to be in the range between 0.3 and 0.4 (Obata et al., 2004; Stephan et al., 2007); here we use $E_0 = 0.34$.

Appendix B. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2015.10.025.

References

Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P., Friston, K.J., 2012. Canonical microcircuits for predictive coding. Neuron 76, 695–711.

Boas, D.A., Jones, S.R., Devor, A., Huppert, T.J., Dale, A.M., 2008. A vascular anatomical network model of the spatio-temporal response to brain activation. NeuroImage 40, 1116–1129.

Buxton, R.B., Wong, E.C., Frank, L.R., 1998. Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. Magn. Reson. Med. 39, 855–864.

Chen, Y., Namburi, P., Elliott, L.T., Heinzle, J., Soon, C.S., Chee, M.W., Haynes, J.D., 2011. Cortical surface-based searchlight decoding. NeuroImage 56, 582–592.

Chen, G., Wang, F., Gore, J.C., Roe, A.W., 2013. Layer-specific BOLD activation in awake monkey V1 revealed by ultra-high spatial resolution functional magnetic resonance imaging. NeuroImage 64, 147–155.

Cheng, K., Waggoner, R.A., Tanaka, K., 2001. Human ocular dominance columns as revealed by high-field functional magnetic resonance imaging. Neuron 32, 359–374.

Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis. I. Segmentation and surface reconstruction. NeuroImage 9, 179–194.

David, O., Kiebel, S.J., Harrison, L.M., Mattout, J., Kilner, J.M., Friston, K.J., 2006. Dynamic causal modeling of evoked responses in EEG and MEG. NeuroImage 30, 1255–1272.

De Martino, F., Zimmermann, J., Muckli, L., Ugurbil, K., Yacoub, E., Goebel, R., 2013. Cortical depth dependent functional responses in humans at 7 T: improved specificity with 3D GRASE. PLoS One 8, e60514.

Donahue, M.J., Hoogduin, H., van Zijl, P.C., Jezzard, P., Luijten, P.R., Hendrikse, J., 2011. Blood oxygenation level-dependent (BOLD) total and extravascular signal changes and DeltaR2* in human visual cortex at 1.5, 3.0 and 7.0 T. NMR Biomed. 24, 25–34.

Douglas, R.J., Martin, K.A., 2004. Neuronal circuits of the neocortex. Annu. Rev. Neurosci. 27, 419–451.

Douglas, R.J., Martin, K.A., 2007. Mapping the matrix: the ways of neocortex. Neuron 56, 226–238.

Douglas, R.J., Martin, K.A.C., Whitteridge, D., 1989. A canonical microcircuit for neocortex. Neural Comput. 1, 480–488.

Dumoulin, S.O., Wandell, B.A., 2008. Population receptive field estimates in human visual cortex. NeuroImage 39, 647–660.

Duvernoy, H.M., Delon, S., Vannson, J.L., 1981. Cortical blood vessels of the human brain. Brain Res. Bull. 7, 519–579.

Feinberg, D.A., Moeller, S., Smith, S.M., Auerbach, E., Ramanna, S., Gunther, M., Glasser, M.F., Miller, K.L., Ugurbil, K., Yacoub, E., 2010. Multiplexed echo planar imaging for sub-second whole brain FMRI and fast diffusion imaging. PLoS One 5, e15710.

Fellemann, D.J., van Essen, D.C., 1991. Distributed hierarchical processing in the primate cerebral cortex. Cereb. Cortex 1, 1–47.

Fischl, B., Dale, A.M., 2000. Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proc. Natl. Acad. Sci. U. S. A. 97, 11050–11055.

Fischl, B., Sereno, M.I., Dale, A.M., 1999. Cortical surface-based analysis. II: inflation, flattening, and a surface-based coordinate system. NeuroImage 9, 195–207.

Friston, K.J., 2002. Bayesian estimation of dynamical systems: an application to fMRI. NeuroImage 16, 513–530.

Friston, K., 2005. A theory of cortical responses. Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci. 360, 815–836.

Friston, K.J., Mechelli, A., Turner, R., Price, C.J., 2000. Nonlinear responses in fMRI: the Balloon model, Volterra kernels, and other hemodynamics. NeuroImage 12, 466–477.

Friston, K.J., Harrison, L., Penny, W., 2003. Dynamic causal modelling. NeuroImage 19, 1273–1302.

Goense, J.B., Logothetis, N.K., 2006. Laminar specificity in monkey V1 using high-resolution SE-fMRI. Magn. Reson. Imaging 24, 381–392.

Goense, J., Merkle, H., Logothetis, N.K., 2012. High-resolution fMRI reveals laminar differences in neurovascular coupling between positive and negative BOLD responses. Neuron 76, 629–639.

Havlicek, M., Roebroeck, A., Friston, K., Gardumi, A., Ivanov, D., Uludag, K., 2015. Physiologically informed dynamic causal modeling of fMRI data. NeuroImage 122, 355–372.

Heidemann, R.M., Ivanov, D., Trampel, R., Fasano, F., Meyer, H., Pfeuffer, J., Turner, R., 2012. Isotropic submillimeter fMRI in the human brain at 7 T: combining reduced field-of-view imaging and partially parallel acquisitions. Magn. Reson. Med. 68, 1506–1516.

Heinzle, J., Hepp, K., Martin, K.A., 2007. A microcircuit model of the frontal eye fields. J. Neurosci. 27, 9341–9353.

Herman, P., Sanganahalli, B.G., Blumenfeld, H., Rothman, D.L., Hyder, F., 2013. Quantitative basis for neuroimaging of cortical laminae with calibrated functional MRI. Proc. Natl. Acad. Sci. U. S. A. 110, 15115–15120.

Hirano, Y., Stefanovic, B., Silva, A.C., 2011. Spatiotemporal evolution of the functional magnetic resonance imaging response to ultrashort stimuli. J. Neurosci. 31, 1440–1447.

Huber, L., Goense, J., Kennerley, A.J., Ivanov, D., Krieger, S.N., Lepsien, J., Trampel, R., Turner, R., Moller, H.E., 2014. Investigation of the neurovascular coupling in positive and negative BOLD responses in human brain at 7 T. NeuroImage 97, 349–362.

Huber, L., Goense, J., Kennerley, A.J., Trampel, R., Guidi, M., Reimer, E., Ivanov, D., Neef, N., Gauthier, C.J., Turner, R., Moller, H.E., 2015. Cortical lamina-dependent blood volume changes in human brain at 7 T. NeuroImage 107, 23–33.

Ivanov, D., Schäfer, A., Deistung, A., Streicher, M.N., Kabisch, S., Henseler, I., Roggenhofer, E., Jochimsen, T.H., Ferdinand, Schweser F., Reichenbach, J.R., Uludag, K., Turner, R., 2013. In vivo estimation of the transverse relaxation time dependence of blood on oxygenation level at 7 Tesla. Proceedings of the International Society of Magnetic Resonance in Medicine 21. Salt Lake City, Utah, USA, p. 2472.

Jin, T., Kim, S.G., 2008. Cortical layer-dependent dynamic blood oxygenation, cerebral blood flow and cerebral blood volume responses during visual stimulation. NeuroImage 43, 1–9.

Kim, T., Kim, S.-G., 2011. Temporal dynamics and spatial specificity of arterial and venous blood volume changes during visual stimulation: implication for BOLD quantification. J. Cereb. Blood Flow Metab. 31, 1211–1222.

Koopmans, P.J., Barth, M., Norris, D.G., 2010a. Layer-specific BOLD activation in human V1. Hum. Brain Mapp. 31, 1297–1304.

Koopmans, P.J., Visser, E., Norris, D.G., Barth, M., 2010b. Layer-specific differential activation in human V1 at 3 T Using 3D EPI. 18Proceedings of the International Society of Magnetic Resonance in Medicine 18. Stockholm, Sweden, p. 3446.

Koopmans, P.J., Barth, M., Orzada, S., Norris, D.G., 2011. Multi-echo fMRI of the cortical laminae in humans at 7 T. NeuroImage 56, 1276–1285.

Kriegeskorte, N., Cusack, R., Bandettini, P., 2010. How does an fMRI voxel sample the neuronal activity pattern: compact-kernel or complex spatiotemporal filter? NeuroImage 49, 1965–1976.

Mandeville, J.B., Marota, J.J., Ayata, C., Zaharchuk, G., Moskowitz, M.A., Rosen, B.R., Weisskoff, R.M., 1999. Evidence of a cerebrovascular postarteriole Windkessel with delayed compliance. J. Cereb. Blood Flow Metab. 19, 679–689.

Markov, N.T., Vezoli, J., Chameau, P., Falchier, A., Quilodran, R., Huissoud, C., Lamy, C., Misery, P., Giroud, P., Ullman, S., Barone, P., Dehay, C., Knoblauch, K., Kennedy, H., 2014. Anatomy of hierarchy: feedforward and feedback pathways in macaque visual cortex. J. Comp. Neurol. 522, 225–259.

Moeller, S., Yacoub, E., Olman, C.A., Auerbach, E., Strupp, J., Harel, N., Ugurbil, K., 2010. Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI. Magn. Reson. Med. 63, 1144–1153.

Moran, R.J., Jung, F., Kumagai, T., Endepols, H., Graf, R., Dolan, R.J., Friston, K.J., Stephan, K.E., Tittgemeyer, M., 2011. Dynamic causal models and physiological inference: a validation study using isoflurane anaesthesia in rodents. PLoS One 6, e22790.

Moran, R., Pinotsis, D.A., Friston, K., 2013. Neural masses and fields in dynamic causal modeling. Front. Comput. Neurosci. 7, 57.

Obata, T., Liu, T.T., Miller, K.L., Luh, W.-M., Wong, E.C., Frank, L.R., Buxton, R.B., 2004. Discrepancies between BOLD and flow dynamics in primary and supplementary motor areas: application of the balloon model to the interpretation of BOLD transients. NeuroImage 21, 144–153.

Olman, C.A., Harel, N., Feinberg, D.A., He, S., Zhang, P., Ugurbil, K., Yacoub, E., 2012. Layer-specific fMRI reflects different neuronal computations at different depths in human V1. PLoS One 7, e32536.

Parkes, L.M., Schwarzbach, J.V., Bouts, A.A., Deckers, R.H., Pullens, P., Kerskens, C.M., Norris, D.G., 2005. Quantifying the spatial resolution of the gradient echo and spin echo BOLD response at 3 Tesla. Magn. Reson. Med. 54, 1465–1472.

Penny, W.D., 2012. Comparing dynamic causal models using AIC, BIC and free energy. NeuroImage 59, 319–330.

Penny, W.D., Stephan, K.E., Mechelli, A., Friston, K.J., 2004. Comparing dynamic causal models. NeuroImage 22, 1157–1172.

Penny, W.D., Stephan, K.E., Daunizeau, J., Rosa, M.J., Friston, K.J., Schofield, T.M., Leff, A.P., 2010. Comparing families of dynamic causal models. PLoS Comput. Biol. 6, e1000709.

Polimeni, J.R., Fischl, B., Greve, D.N., Wald, L.L., 2010. Laminar analysis of 7 T BOLD using an imposed spatial activation pattern in human V1. NeuroImage 52, 1334–1346.

Poser, B.A., Koopmans, P.J., Witzel, T., Wald, L.L., Barth, M., 2010. Three dimensional echo-planar imaging at 7 Tesla. NeuroImage 51, 261–266.

Rao, R.P., Ballard, D.H., 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. 2, 79–87.

Reichold, J., Stampanoni, M., Lena Keller, A., Buck, A., Jenny, P., Weber, B., 2009. Vascular graph model to simulate the cerebral blood flow in realistic vascular networks. J. Cereb. Blood Flow Metab. 29, 1429–1443.

Santisakultarm, T.P., Cornelius, N.R., Nishimura, N., Schafer, A.I., Silver, R.T., Doerschuk, P.C., Olbricht, W.L., Schaffer, C.B., 2012. In vivo two-photon excited fluorescence microscopy reveals cardiac- and respiration-dependent pulsatile blood flow in cortical blood vessels in mice. Am. J. Physiol. Heart Circ. Physiol. 302, H1367–H1377.

Siero, J.C., Petridou, N., Hoogduin, H., Luijten, P.R., Ramsey, N.F., 2011. Cortical depth-dependent temporal dynamics of the BOLD response in the human brain. J. Cereb. Blood Flow Metab. 31, 1999–2008.

Siero, J.C., Ramsey, N.F., Hoogduin, H., Klomp, D.W., Luijten, P.R., Petridou, N., 2013. BOLD specificity and dynamics evaluated in humans at 7 T: comparing gradient-echo and spin-echo hemodynamic responses. PLoS One 8, e54560.

Siero, J.C., Hendrikse, J., Hoogduin, H., Petridou, N., Luijten, P., Donahue, M.J., 2015. Cortical depth dependence of the BOLD initial dip and poststimulus undershoot in human visual cortex at 7 Tesla. Magn. Reson. Med. 73, 2283–2295.

Silvennoinen, M.J., Clingman, C.S., Golay, X., Kauppinen, R.A., van Zijl, P.C., 2003. Comparison of the dependence of blood R2 and R2* on oxygen saturation at 1.5 and 4.7 Tesla. Magn. Reson. Med. 49, 47–60.

Smith, F.W., Muckli, L., 2010. Nonstimulated early visual areas carry information about surrounding context. Proc. Natl. Acad. Sci. U. S. A. 107, 20099–20103.

Stephan, K.E., Baldeweg, T., Friston, K.J., 2006. Synaptic plasticity and dysconnection in schizophrenia. Biol. Psychiatry 15, 929–939.

Stephan, K.E., Weiskopf, N., Drysdale, P.M., Robinson, P.A., Friston, K.J., 2007. Comparing hemodynamic models with DCM. NeuroImage 38, 387–401.

Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., Friston, K.J., 2009. Bayesian model selection for group studies. NeuroImage 46, 1004–1017.

Tian, P., Teng, I.C., May, L.D., Kurz, R., Lu, K., Scadeng, M., Hillman, E.M., De Crespigny, A.J., D'Arceuil, H.E., Mandeville, J.B., Marota, J.J., Rosen, B.R., Liu, T.T., Boas, D.A., Buxton, R.B., Dale, A.M., Devor, A., 2010. Cortical depth-specific microvascular dilation underlies laminar differences in blood oxygenation level-dependent functional MRI signal. Proc. Natl. Acad. Sci. U. S. A. 107, 15246–15251.

Turner, R., 2002. How much cortex can a vein drain? Downstream dilution of activation-related cerebral blood oxygenation changes. NeuroImage 16, 1062–1067.

Uludag, K., Muller-Bierl, B., Ugurbil, K., 2009. An integrative model for neuronal activity-induced signal changes for gradient and spin echo functional imaging. NeuroImage 48, 150–165.

Waehnert, M.D., Dinse, J., Weiss, M., Streicher, M.N., Waehnert, P., Geyer, S., Turner, R., Bazin, P.L., 2014. Anatomically motivated modeling of cortical laminae. NeuroImage 93, 210–220.

Weber, B., Keller, A.L., Reichold, J., Logothetis, N.K., 2008. The microvascular system of the striate and extrastriate visual cortex of the macaque. Cereb. Cortex 18, 2318–2330.

Welvaert, M., Rosseel, Y., 2013. On the definition of signal-to-noise ratio and contrast-to-noise ratio for fMRI data. PLoS One 8, e77089.

Yacoub, E., Shmuel, A., Logothetis, N., Ugurbil, K., 2007. Robust detection of ocular dominance columns in humans using Hahn Spin Echo BOLD functional MRI at 7 Tesla. NeuroImage 37, 1161–1177.

Yacoub, E., Harel, N., Ugurbil, K., 2008. High-field fMRI unveils orientation columns in humans. Proc. Natl. Acad. Sci. U. S. A. 105, 10607–10612.

Yu, X., Glen, D., Wang, S., Dodd, S., Hirano, Y., Saad, Z., Reynolds, R., Silva, A.C., Koretsky, A.P., 2012. Direct imaging of macrovascular and microvascular contributions to BOLD fMRI in layers IV–V of the rat whisker–barrel cortex. NeuroImage 59, 1451–1460.

Yu, X., Qian, C., Chen, D.Y., Dodd, S.J., Koretsky, A.P., 2014. Deciphering laminar-specific neural inputs with line-scanning fMRI. Nat. Methods 11, 55–58.

Zhao, F., Wang, P., Hendrich, K., Ugurbil, K., Kim, S.G., 2006. Cortical layer-dependent BOLD and CBV responses measured by spin-echo and gradient-echo fMRI: insights into hemodynamic regulation. NeuroImage 30, 1149–1160.

Zhao, J.M., Clingman, C.S., Narvainen, M.J., Kauppinen, R.A., van Zijl, P.C., 2007. Oxygenation and hematocrit dependence of transverse relaxation rates of blood at 3 T. Magn. Reson. Med. 58, 592–597.