Haemodynamic modelling

Glaser DE, Friston KJ, Mechelli A, Turner R and Price CJ

1. Introduction

There is a growing appreciation of the importance of nonlinearities in evoked responses in fMRI, particularly with the advent of event-related fMRI. These nonlinearities are commonly expressed as interactions among stimuli that can lead to the suppression and increased latency of responses to a stimulus that are incurred by a preceding stimulus. We have presented previously a model-free characterisation of these effects using generic techniques from nonlinear system identification, namely a Volterra series formulation. At the same time Buxton *et al* (1998) described a plausible and compelling dynamical model of haemodynamic signal transduction in fMRI. Subsequent work by Mandeville *et al* (1999) provided important theoretical and empirical constraints on the form of the dynamic relationship between blood flow and volume that underpins the evolution of the fMRI signal. In this chapter we combine these system identification and model-based approaches and ask whether the Balloon model is sufficient to account for the nonlinear behaviours observed in real time series. We conclude that it can, and furthermore the model parameters that ensue are biologically plausible. This conclusion is based on the observation that the Balloon model can produce Volterra kernels that emulate empirical kernels.

To enable this evaluation we have had to embed the Balloon model in a haemodynamic input-state-output model that included the dynamics of perfusion changes that are contingent on underlying synaptic activation. This chapter presents (i) the full haemodynamic model (ii), how its associated Volterra kernels can be derived and (iii) addresses the model's validity in relation to empirical nonlinear characterisations of evoked responses in fMRI and other neurophysiological constraints.

2. Background

This chapter is about modelling the relationship between neural activity and the BOLD (blood oxygenation level dependent) fMRI signal. Before describing a comprehensive model which can account for the most important types of non-linearity empirically observed from fMRI studies, it is worth briefly putting this work into its proper context. Essentially there are three things that need to be modelled in order to understand the neural-BOLD relationship. We must be clear about which aspects of neural activity are of interest to us, and also which give rise to the signals we measure. We should clarify the nature and properties of the mechanisms relating this activity, through metabolic demand, to the blood supply to the tissue containing the neurons. Finally, we need a model of how these changes in blood supply affect the signal measured in the scanner.

That there is a connection between blood supply and brain activity has been known for over 100 years. In their seminal paper, Roy and Sherrington (1890) concluded that functional activity increased blood flow and inferred that there was a coupling generating increased blood flow in response to increased metabolic demand. Interestingly, their observation of the consequences of metabolic demand came before the demonstration of the increase in demand itself. It was more than seventy years later that the regional measurement of the metabolic changes was convincingly achieved using an autoradiographic technique which used a substitute for glucose, called deoxyglucose (2DG) radioactively labelled with C14. 2DG enters the cells by the same mechanisms as glucose but is not metabolized and thus accumulates inside the cells at a rate which is dependent on their metabolic activity. By examining the density of labelled 2DG in brain slices, Sokoloff and colleagues (Kennedy *et al*, 1976) obtained functional maps of the activity during the period in which 2DG was injected. This activity period was generally around 45 mins, which limited the time resolution of the technique. In addition, only one measurement per subject could be made since the technique involves the sacrifice of the animal (further developments allowed the injection of two tracers, but this was still very restrictive). However the spatial resolution could be microscopic since the label is contained in the cells themselves rather than being limited to the blood vessels surrounding them. Through theoretical modelling of the enzyme kinetics for the uptake of 2DG and practical experiments, the relationships between neural function and glucose

metabolism have been established and underpin the development of "metabolic encephalography".

Positron emission tomography (PET) measures an intermediate stage in the chain linking neural activity via metabolism to the BOLD signal. By using a tracer such as O15 labelled water, one can measure changes in regional cerebral blood flow (rCBF) which accompany changes in neural activity. This was originally thought of as an autoradiographic technique, but has many advantages over 2DG and is clearly much less invasive making it suitable for human studies. Also, substantially shorter times are required for measurements, typically well below a minute. As suggested above, the elucidation of the mechanisms underlying the coupling of neural activity and blood flow lags behind the exploitation of the phenomenon. There are several candidate signals including the diffusible second messengers such as nitric oxide or intravascular responses to changes in blood oxygenation level caused by changes in oxygen consumption, and this remains an active area of research independently of its consequences for models of functional brain imaging.

In the treatment below, we follow evidence from Miller *et al* (2000) among others, and assume that blood flow and neural activity are linearly related over normal ranges. However, there are ongoing arguments about the nature of the linkage between neural activity, the rate of metabolism of oxygen and cerebral blood flow. Some PET studies have suggested that while an increase in neural activity produces a proportionate increase in glucose metabolism and cerebral blood flow, oxygen consumption does not increase proportionately (PT Fox and ME Raichle, 1986). This decoupling between blood flow and oxidative metabolism is known as the 'anaerobic brain' hypothesis by analogy with muscle physiology. Arguing against this position, other groups have adopted an even more radical interpretation. They suggest that immediately following neural stimulation there is a transient decoupling between neural activity and blood flow (Vanzetta and Grinvald, 1999). By this argument, there is an immediate increase in oxidative metabolism which produces a transient localized increase in deoxyhaemoglobin. Only later do the mechanisms regulating blood flow kick in causing the observed increase in rCBF and hence blood volume. Evidence for this position comes from optical imaging studies and depends on modelling the absorption and light-scattering properties of

cortical tissue and the relevant chromophores, principally (de)oxyhaemoglobin. Other groups have questioned these aspects of the work, and the issue remains controversial (Lindauer *et al* 2001). One possible consequence of this position is that better spatial resolution would be obtained by focussing on this early phase of the haemodynamic response.

As this chapter will demonstrate, the situation is even more complicated with regard to functional magnetic resonance imaging (fMRI) using a blood oxygenation level dependent (BOLD) contrast. As its name suggests, the technique exploits the amount of oxygen in the blood as a marker for neural activity, exploiting the fact that deoxyhaemoglobin is less diamagnetic than oxyhaemoglobin. Blood oxygenation level refers to the *proportion* of oxygenated blood but the signal depends on the total amount of deoxyhaemoglobin and so the total volume of blood is a factor. Another factor is the change in the amount of oxygen leaving the blood to enter the tissue and meet changes in metabolic demand. Since the blood which flows into the capillary bed is fully oxygenated, changes in blood flow also change blood oxygenation level. Finally the elasticity of the vascular tissue of the veins and venules means that an increase in blood flow causes an increase in blood volume. All these factors are modelled and discussed in the body of the chapter. Of course even more factors can be considered; for example Mayhew and colleagues (Zheng *et al*, 2002) have extended the treatment described here to include (among others) the dynamics of oxygen buffered in the tissue.

Notwithstanding these complications it is a standard assumption that "the fMRI signal is approximately proportional to a measure of local neural activity" (reviewed in Heeger & Ress, 2002), and this linear model is still used in many studies particularly where interstimulus intervals are more than a second or two. Empirical evidence against this hypothesis is outlined below, but note that there are now theoretical objections too. In particular, the models which have been developed to account for observed non-linearities embody our best knowledge about the physiological mechanisms at work in the regulation of blood volume and oxygenation. Since they generate non-linearities in BOLD response given reasonable choices for the parameters (discussed below), one might consider the genie to have been let out of the bottle.

The last link in the chain concerns the relation between a complete description of the relevant aspects of blood supply and the physics underlying the BOLD signal. While this is not the principal focus of this chapter, a couple of simple points are worth emphasising. Firstly, differently sized blood vessels will give different changes in BOLD signal for the same changes in blood flow, volume and oxygenation. This is because of differences in the inhomogeneity of the magnetic fields in their vicinity. Secondly, and for partially related reasons, heuristic equations as employed in this and other models are dependent on the strength of the magnet. In particular the equation used here may be relevant only for 1.5 T scanners, although other versions for different field strength have been developed.

Finally a word about "neural activity". So far in the discussion we have deliberately not specified what type of neural activity we are considering. Here again there are theoretical and practical issues. Firstly it is worth remembering that different electrophysiological measures can emphasise different elements of neural firing (also see below). In particular, recording of multiple single units with an intracortical microelectrode can tend to sample action potentials from large pyramidal output neurons. Such studies are frequently referred to when characterising the response properties of a primate cortical area. However, consideration of the metabolic demands of various cellular processes suggests that spiking is not the major drain on the resources of a cell but rather that synaptic transmission and conductances of post-synaptic potentials as well as cytoskeletal turnover are the dominating forces. Of course such processes are just as important, whether in interneurons and whether excitatory or inhibitory. An example of the difference between these two views of the cortex might be feed-forward vs. feed-back activity in low-level visual cortex. Indeed BOLD fMRI experiments in humans have shown good agreement with studies of spiking in V1 in response to modulating the contrast of a visual stimulus, but attentional (top-down) modulation effects in V1 have proved elusive in monkey electrophysiological studies but robust with BOLD studies in humans. Aside from their neurobiological significance, such discrepancies must be born in mind when defining the neural activity to which the BOLD signal might be linearly responding.

A further subtlety relates to the modelled time course of the neural activity. Even in everyday analysis of functional imaging data it is natural to separate the model of the

response into neural and haemodynamic components. However, a typical set of spikes or block functions often used to model the neural activity will fail to capture adaptation and response transients which should be well known from the neurophysiological literature. (Note that a recent set of studies has deliberately exploited these effects (Grill-Spector & Malach, 2001)). In the worst case an elaborate model designed to capture non-linearities in the BOLD response may inadvertently pick up such components of the neural response, and in any case careful stimulus design and modelling of neural responses is called for.

Recent studies using simultaneous fMRI and intracortical electrical recording in monkey have empirically validated many of the theoretical points considered above (Logothetis *et al*, 2001). In particular, the closeness of the BOLD signal to LFP and MUA rather than spiking activity have been emphasised. These studies also demonstrated that the linear assumption can predict up to 90% of the variance in BOLD responses in some cortical regions. However there was considerable variability in the accuracy of prediction, with nearby sites sometimes being substantially worse. Overall, substantial non-linearities were observed between stimulus contrast, blood flow and BOLD signals.

Having surveyed the general issues surrounding the coupling of neural activity and the BOLD signal, we will now proceed to outline a specific and detailed model. This should be considered as a partial instantiation of current knowledge and further extensions of this model have been proposed which incoporate new data. Mechelli (this volume) also presents further empirical verification of the parameter regime proposed here. What follows is largely a reprise of Friston et al (2000), and contains some advanced mathematical material

3. Nonlinear evoked responses

We now focus on the nonlinear aspects of evoked responses in functional neuroimaging and presents a dynamical approach to modelling and characterising event-related signals in fMRI. We aim to: (i) show that the Balloon/Windkessel model (Buxton and Frank 1997, Buxton *et al* 1998, Mandeville *et al* 1999) is sufficient to account for nonlinearities in event-related responses that are seen empirically and (ii) describe a nonlinear dynamical model that couples changes in synaptic activity to fMRI signals. This haemodynamic model obtains by combining the Balloon/Windkessel model (henceforth Balloon model) with a model of how synaptic activity causes changes in regional flow.

In Friston *et al* (1994) we presented a linear model of haemodynamic responses in fMRI time-series, wherein underlying neuronal activity (inferred on the basis of changing stimulus or task conditions) is convolved, or smoothed with a *haemodynamic response function*. In Friston *et al* (1998) we extended this model to cover nonlinear responses using a Volterra series expansion. At the same time Buxton and colleagues developed a mechanistically compelling model of how evoked changes in blood flow were transformed into a blood oxygenation level dependent (BOLD) signal (Buxton *et al* 1998). A component of the Balloon model, namely the relationship between blood flow and volume, was then elaborated in the context of standard windkessel theory by Mandeville *et al* (1999). The Volterra approach, in contradistinction to other nonlinear characterisation of haemodynamic responses (*c.f.* Vazquez and Noll 1996), is model-independent, in the sense that Volterra series can model the behaviour of any nonlinear time-invariant dynamical system[1]. The principal aim of this work was to see if the theoretically motivated Balloon model would be sufficient to explain the nonlinearities embodied in a purely empirical Volterra characterisation.

[1] In principle Volterra series can represent any dynamical input-state-output system and in this sense a characterisation in terms of Volterra kernels is model independent. However, by using basis functions to constrain the solution space, constraints are imposed on the form of the kernels and, implicitly, the underlying dynamical system (i.e. state-space representation). The characterisation is therefore only assumption free to the extent the basis set is sufficiently comprehensive.

3.1. Volterra Series

Volterra series express the output of a system, in this case the BOLD signal from a particular voxel, as a function of some input, here the assumed synaptic activity that is changed experimentally. This series is a function of the input over its recent history and is expressed in terms of generalised convolution kernels. Volterra series are often referred to as nonlinear convolutions or polynomial expansions with memory. They are simply Taylor expansions extended to cover dynamical input-state-output systems by considering the effect of the input now and at all times in the recent past. The zeroth order kernel is simply a constant about which the response varies. The first order kernel represents the weighting applied to a sum of inputs over the recent past (*c.f.* the haemodynamic response function) and can be thought of as the change in output for a change in the input at each time point. Similarly, the second order coefficients represent interactions that are simply the effect of the input at one point in time on its contribution at another. The second order kernel comprises coefficients that are applied to interactions among (i.e. products of) inputs, at different times in the past, to predict the response.

In short the output can be considered a nonlinear convolution of the input where nonlinear behaviours are captured by high order kernels. For example the presence of a stimulus can be shown to attenuate the magnitude of, and induce a longer latency in, the response to a second stimulus that occurs within a second or so. The example shown in Figure 1 comes from our previous analysis (Friston *et al* 1998) and shows how a preceding stimulus can modify the response to a subsequent stimulus. This sort of effect led to the notion of *haemodynamic refractoriness* and is an important example of nonlinearity in fMRI time-series.

The important thing about Volterra series is that they do not refer to all the hidden state variables that mediate between the input and output (e.g. blood flow, venous volume, oxygenation, the dynamics of endothelium derived relaxing factor, kinetics of cerebral metabolism *etc*.). This renders them very powerful because they provide for a complete specification of the dynamical behaviour of a system without ever having to measure the state variables or make any assumptions about how these variables interact to produce a response. On the other hand the Volterra formulation is impoverished because it yields no mechanistic insight into how the response is mediated. The alternative is to posit

some model of interacting state variables and establish the validity of that model in relation to observed input-output behaviours and the dynamics of the state variables themselves. This involves specifying a series of differential equations that express the change in one state variable as a function of the others and the input. Once these equations are specified the equivalent Volterra representation can be derived analytically (see the Appendix for details). The Balloon model is a comprehensive example of such a model.

3.2. The Balloon model

The Balloon model (Buxton and Frank 1997, Buxton *et al* 1998) is an input-state-output model with two state variables: volume (v) and deoxyhaemoglobin content (q). The input to the system is blood flow (f_{in}) and the output is the BOLD signal (y). The BOLD signal is partitioned into an extra and intra-vascular component, weighted by their respective volumes. These signal components depend on the deoxyhaemoglobin content and render the signal a nonlinear function of v and q. The effect of flow on v and q (see below) determines the output and it is these effects that are the essence of the Balloon model: Increases in flow effectively inflate a venous 'balloon' such that deoxygenated blood is diluted and expelled at a greater rate. The clearance of deoxyhaemoglobin reduces intra-voxel dephasing and engenders an increase in signal. Before the balloon has inflated sufficiently the expulsion and dilution may be insufficient to counteract the increased delivery of deoxygenated blood to the venous compartment and an 'early dip' in signal may be expressed. After the flow has peaked, and the balloon has relaxed again, reduced clearance and dilution contribute to the post-stimulus undershoot commonly observed. This is a simple and plausible model that is predicated on a minimal set of assumptions and relates closely to the windkessel formulation of Mandeville *et al* (1999). Furthermore the predictions of the Balloon model concur with the steady-state models of Hoge and colleagues, and their elegant studies of the relationship between blood flow and oxygen consumption in human visual cortex (e.g. Hoge *et al* 1999).

The Balloon model is inherently nonlinear and may account for the sorts of nonlinear interactions revealed by the Volterra formulation. One simple test of this hypothesis is to see if the Volterra kernels associated with the Balloon model compare with those derived empirically. The Volterra kernels estimated in Friston *et al* (1998) clearly did not use

flow as input because flow is not measurable with BOLD fMRI. The input comprised a stimulus function as an index of synaptic activity. In order to evaluate the Balloon model in terms of these Volterra kernels it has to be extended to accommodate the dynamics of how flow is coupled to synaptic activity encoded in the stimulus function. This chapter presents one such extension.

In summary the Balloon model deals with the link between flow and BOLD signal. By extending the model to cover the dynamic coupling of synaptic activity and flow a complete model, relating experimentally-induced changes in neuronal activity to BOLD signal, obtains. The input-output behaviour of this model can be compared to the real brain in terms of their respective Volterra kernels.

The remainder of this chapter is divided into three sections. In the next section we present a haemodynamic model of the coupling between synaptic activity and BOLD response that builds upon the Balloon model. The second section presents an empirical evaluation of this model by comparing its Volterra kernels with those obtained using real fMRI data. This is not a trivial exercise because; (i) there is no guarantee that the Balloon model could produce the complicated forms of the kernels seen empirically and, (ii) even if it could, the parameters needed to do so may be biologically implausible. This section provides estimates of these parameters, which allow some comment on the face validity of the model, in relation to known physiology. The final section presents a discussion of the results in relation to known biophysics and neurophysiology.

This chapter is concerned with the validation and evaluation of the Balloon model, in relation to the Volterra characterisations, and the haemodynamic model presented below in relation to real haemodynamics. Subsequent papers will use the model to address some important issues related to experimental design and the sorts of neuronal dynamics that BOLD signals are most sensitive to.

4. The haemodynamic model

In this section we describe a haemodynamic model that mediates between synaptic activity and measured BOLD responses. This model essentially combines the Balloon model and a simple linear dynamical model of changes in regional cerebral blood flow (rCBF) caused by neuronal activity. The model architecture is summarised in Figure 2.

To motivate the model components more clearly we will start at the output and work towards the input.

4.1. The Balloon component

This component links rCBF and the BOLD signal as described in Buxton *et al* (1998). All variables are expressed in normalised form, relative to resting values. The BOLD signal $y(t) = \lambda(v, q, E_0)$ is taken to be a static nonlinear function of normalised venous volume (v), normalised total deoxyhaemoglobin voxel content (q) and resting net oxygen extraction fraction by the capillary bed (E_0)

$$\begin{aligned} y(t) &= \lambda(v, q, E_0) = V_0\left(k_1(1-q) + k_2(1-q/v) + k_3(1-v)\right) \\ k_1 &= 7E_0 \\ k_2 &= 2 \\ k_3 &= 2E_0 - 0.2 \end{aligned} \qquad 1$$

where V_0 is resting blood volume fraction. This signal comprises a volume-weighted sum of extra- and intra-vascular signals that are functions of volume and deoxyhaemoglobin content. The latter are the state variables whose dynamics need specifying. The rate of change of volume is simply

$$\tau_0 \dot{v} = f_{in} - f_{out}(v) \qquad 2$$

See Mandeville *et al* (1999) for an excellent discussion of this equation in relation to windkessel theory. Eq(2) says that volume changes reflect the difference between inflow f_{in} and outflow f_{out} from the venous compartment with a time constant τ_0. This constant represents the mean transit time (*i.e.* the average time it takes to traverse the venous compartment or for that compartment to be replenished) and is V_0 / F_0 where F_0 is resting flow. The physiology of the relationship between flow and volume is determined by the evolution of the transit time. Mandeville *et al* (1999) reformulated the temporal evolution of transit time into a description of the dynamics of resistance and capacitance of the balloon using windkessel theory ('windkessel' means leather bag).

This enabled them to posit a form for the temporal evolution of a downstream elastic response to arteriolar vasomotor changes and estimate mean transit times using measurements of volume and flow, in rats, using fMRI and laser-Doppler flowmetry. We will compare these estimates to our empirical estimates in the next section.

Note that outflow is a function of volume. This function models the balloon-like capacity of the venous compartment to expel blood at a greater rate when distended. We model it with a single parameter α based on the windkessel model

$$f_{out}(v) = v^{1/\alpha} \qquad 3$$

where $1/\alpha = \gamma + \beta$. (c.f. Eq(6) in Mandeville *et al* 1999). $\gamma = 2$ represents laminar flow. $\beta > 1$ models diminished volume reserve at high pressures and can be thought of as the ratio of the balloon's capacitance to its compliance. At steady state empirical results from PET suggest $\alpha \approx 0.38$ (Grubb *et al* 1974). However, when flow and volume are changing dynamically, this value is smaller. Mandeville *et al* (1999) were the first to measure the dynamic flow-volume relationship and estimated $\alpha \approx 0.18$, after 6 seconds of stimulation, with a projected asymptotic [steady-state] value of 0.36.

The change in deoxyhaemoglobin $\dot{q}$ reflects the delivery of deoxyhaemoglobin into the venous compartment minus that expelled (outflow times concentration)

$$\tau_0 \dot{q} = f_{in} \frac{E(f_{in}, E_0)}{E_0} - f_{out}(v) q / v \qquad 4$$

where $E(f_{in}, E_0)$ is the fraction of oxygen extracted from the inflowing blood. This is assumed to depend on oxygen delivery and is consequently flow-dependent. A reasonable approximation for a wide range of transport conditions is (Buxton *et al* 1998)

$$E(f_{in}, E_0) = 1 - (1 - E_0)^{1/f_{in}} \qquad 5$$

The second term in Eq(4) represents an important nonlinearity: The effect of flow on signal is largely determined by the inflation of the balloon, resulting in an increase of $f_{out}(v)$ and clearance of deoxyhaemoglobin. This effect depends upon the concentration of deoxyhaemoglobin such that the clearance attained by the outflow will be severely attenuated when the concentration is low (*e.g.* during the peak response to a prior stimulus). The implications of this will be illustrated in the next section.

This concludes the Balloon model component, where there are only three unknown parameters that determine the dynamics E_0, τ_0 and α, namely resting oxygen extraction fraction (E_0), mean transit time (τ_0) and a stiffness exponent (α) specifying the flow-volume relationship of the venous balloon. The only thing required, to specify the BOLD response, is inflow.

4.2. rCBF component

It is generally accepted that, over normal ranges, blood flow and synaptic activity are linearly related. A recent empirical verification of this assumption can be found in Miller *et al* (2000) who used MRI perfusion imaging to address this issue in visual and motor cortices. After modelling neuronal adaptation they were able to conclude, "Both rCBF responses are consistent with a linear transformation of a simple nonlinear neural response model". Furthermore our own work using PET and fMRI replications of the same experiments suggests that the observed nonlinearities enter into the translation of rCBF into a BOLD response (as opposed to a nonlinear relationship between synaptic activity and rCBF) in the auditory cortices (see Friston *et al* 1998). Under the constraint that the dynamical system linking synaptic activity and rCBF is linear we have chosen the most parsimonious model

$$\dot{f}_{in} = s \qquad 6$$

where *s* is some flow inducing signal defined, operationally, in units corresponding to the rate of change of normalised flow (i.e. sec^{-1}). Although it may seem more natural to express the effect of this signal directly on vascular resistance (*r*), for example $\dot{r} = -s$,

Eq(6) has the more plausible form. This is because the effect of signal (s) is much smaller when r is small (when the arterioles are fully dilated signals such as endothelium-derived relaxing factor or nitric oxide will cause relatively small decrements in resistance). This can be seen by noting Eq(6) is equivalent to $\dot{r} = -r^2 s$, where $f_{in} = 1/r$.

The signal is assumed to subsume many neurogenic and diffusive signal sub-components and is generated by neuronal activity $u(t)$

$$\dot{s} = \varepsilon u(t) - s/\tau_s - (f_{in} - 1)/\tau_f \qquad 7$$

ε, τ_s and τ_f are the three unknown parameters that determine the dynamics of this component of the haemodynamic model. They represent the efficacy with which neuronal activity causes an increase in signal, the time-constant for signal decay or elimination and the time-constant for autoregulatory feedback from blood flow. The existence of this feedback term can be inferred from; (i) post-stimulus undershoots in rCBF (e.g. Irikura *et al* 1994) and (ii) the well-characterised vasomotor signal in optical imaging (Mayhew *et al* 1998). The critical aspect of the latter oscillatory (~0.1 Hz) component of intrinsic signals is that it shows variable phase relationships from region to region, supporting strongly the notion of local closed-loop feedback mechanisms as modelled in Eq(6) and Eq(7).

There are three unknown parameters for each of the two components of the haemodynamic model above (see also Figure 2 for a schematic summary). Figure 3 illustrates the behaviour of the haemodynamic model for typical values of the six parameters ($\varepsilon = 0.5, \tau_s = 0.8, \tau_f = 0.4, \tau_0 = 1, \alpha = 0.2, E_0 = 0.8$ and assuming $V_0 = 0.02$ here and throughout). We have used a very high value for oxygen extraction to accentuate the early dip (see discussion). Following a short-lived neuronal transient a substantial amount of signal is created and starts to decay immediately. This signal induces an increase in flow that itself augments signal decay, to the extent the signal is suppressed below resting levels (see the upper left panel in Figure 3). This behaviour is homologous to a very dampened oscillator. Increases in flow (lower left panel) dilate the

venous balloon which responds by ejecting deoxyhaemoglobin. In the first few hundred milliseconds the net deoxyhaemoglobin (q) increases with an accelerating inflow-dependent delivery. It is then cleared by volume-dependent outflow expressing a negative peak a second or so after the positive volume (v) peak (the broken and solid lines in the upper right panel correspond to q and v respectively). This results in an early dip in the BOLD signal followed by a pronounced positive peak at about 4 seconds (lower right panel) that reflects the combined effects of reduced net deoxyhaemoglobin, increased venous volume and consequent dilution of deoxyhaemoglobin. Note that the rise and peak in volume (solid line in the upper right panel) lags flow by about a second. This is very similar to the predictions of the windkessel formulation and the empirical results presented in Mandeville *et al* (1999) (see their Figure 2). After about 8 seconds the inflow experiences a rebound due to its suppression of the perfusion signal. The reduced venous volume and ensuing outflow permit a re-accumulation of deoxyhaemoglobin and a consequent undershoot in the BOLD signal.

The rCBF component of the haemodynamic model is a linear dynamical system and as such has only zeroth and first order kernels. This means it *cannot* account for the haemodynamic refractoriness and nonlinearities observed in BOLD responses. Although the rCBF component may facilitate the Balloon component's capacity to model nonlinearities (by providing appropriate input), the rCBF component alone cannot generate second order kernels. The question addressed in this chapter is whether the Balloon component can produce second order kernels that are realistic and do so with physiologically plausible parameters.

5. Model parameter estimation

In this section we describe the data used to estimate Volterra kernels. The six unknown parameters of the haemodynamic model that best reproduce these empirical kernels are then identified. By minimising the difference between the model kernels and the empirical kernels the optimal parameters for any voxel can be determined. The critical questions this section addresses are (i) ‘can the haemodynamic model account for the form of empirical kernels up to second order?’ and (ii) ‘are the model parameters required to do this physiologically plausible?’

5.1. Empirical analyses

The data and Volterra kernel estimation are described in detail in Friston *et al* (1998). In brief we obtained fMRI time-series from a single subject at 2 Tesla using a Magnetom VISION (Siemens, Erlangen) whole body MRI system, equipped with a head volume coil. Contiguous multi-slice T_2*-weighted fMRI images were obtained with a gradient echo-planar sequence using an axial slice orientation (TE = 40ms, TR = 1.7 seconds, 64x64x16 voxels). After discarding initial scans (to allow for magnetic saturation effects) each time-series comprised 1200 volume images with 3mm isotropic voxels. The subject listened to monosyllabic or bi-syllabic concrete nouns (i.e. 'dog', 'radio', 'mountain', 'gate') presented at 5 different rates (10 15 30 60 and 90 words per minute) for epochs of 34 seconds, intercalated with periods of rest. The 5 presentation rates were successively repeated according to a Latin Square design.

The data were processed within SPM (Wellcome Department of Cognitive Neurology, http://www.fil.ion.ucl.ac.uk/spm). The time-series were realigned, corrected for movement-related effects and spatially normalised into the standard space of Talairach and Tournoux (1988). The data were smoothed spatially with a 5mm isotropic Gaussian kernel. Volterra kernels were estimated by expanding the kernels in terms of temporal basis functions and estimating the kernel coefficients up to second order using a generalised linear model (Worsley and Friston 1995). The basis set comprised three gamma varieties of increasing dispersion and their temporal derivatives (as described in Friston *et al* 1998).

The stimulus function $u(t)$, the supposed neuronal activity, was simply the word presentation rate at which the scan was acquired. We selected voxels that showed a robust response to stimulation from two superior temporal regions in both hemispheres (see Figure 4). These were the 128 voxels showing the most significant response when testing for the null hypothesis that the first and second order kernels were jointly zero. Selecting these voxels ensured that the kernel estimates had minimal variance.

5.2. Estimating the model parameters

For each voxel we identified the six parameters of the haemodynamic model of the previous section whose kernels corresponded, in a least squares sense, to the empirical kernels for that voxel. To do this we used nonlinear function minimisation as

implemented in MATLAB5 (MathWorks Inc. MA). The model's kernels were computed, for a given parameter vector, as described in the Appendix and entered, with the corresponding empirical estimates, into the objective function that was minimised.

5.3. Results

The model-based and empirical kernels for the first voxel are shown in Figure 5. It can be seen that there is a remarkable agreement both in terms of the first and second order kernels. This is important because it suggests that the nonlinearities inherent in the Balloon component of the haemodynamic model are sufficient to account for the nonlinear responses observed in real time-series. The first order kernel corresponds to the conventional [first order] haemodynamic response function and shows the characteristic peak at about 4 seconds and the post-stimulus undershoot. The empirical undershoot appears more protracted than the model's prediction suggesting that the model is not perfect in every respect. The second order kernel has a pronounced negativity on the upper left, flanked by two smaller positivities. This negativity accounts for the refractoriness seen when two stimuli are temporally proximate, where this proximity is defined by the radius of the negative region. From the perspective of the Balloon model the second stimulus is compromised, in terms of elaborating a BOLD signal, because of the venous pooling, and consequent dilution of deoxyhaemoglobin, incurred by the first stimulus. This means that less deoxyhaemoglobin can be cleared for a given increase in flow. More interesting are the positive regions, which suggest stimuli separated by about 8 seconds should show super-additive effects. This can be attributed to the fact that, during the flow undershoot following the first stimulus, deoxyhaemoglobin concentration is greater than normal (see the upper right panel in Figure 3), thereby facilitating clearance of deoxyhaemoglobin following the second stimulus.

Figure 6 shows the various functions implied by the haemodynamic model parameters averaged over all voxels. These include outflow as a function of venous volume $f_{out}(v,\alpha)$ and oxygen extraction fraction as a function of inflow. The solid line in the upper right panel is extraction *per se* $E(f_{in},E_0)$ and the broken line is the net normalised

delivery of deoxyhaemoglobin to the venous compartment $f_{in}E(f_{in},E_0)/E_0$. Note that although the fraction of oxygen extracted decreases with flow the net delivery of deoxygenated haemoglobin increases with flow. In other words inflow increases *per se* actually reduce signal. It is only the secondary effects of inflow on dilution and volume-dependent outflow that cause an increase in BOLD signal. The lower panel depicts the nonlinear function of volume and deoxyhaemoglobin that represents BOLD signal $y(t)=\lambda(v,q,E_0)$. Here one observes that positive BOLD signals are expressed only when deoxyhaemoglobin is low. The effect of volume is much less marked and tends to affect signal predominantly through dilution. This is consistent with the fact that $k_2 > k_3$ [see Eq(1)] for the value of E_0 estimated for these data.

The distributions of the parameters over voxels are shown in Figure 7 with their mean in brackets at the top of each panel. It should be noted that the data from which these estimates came were not independent. However, given they came from four different brain regions they are remarkably consistent. In the next section we will discuss each of these parameters and the effect it exerts on the BOLD response.

6. Discussion

The main point to be made here is that the Balloon model, suitably extended to incorporate the dynamics of rCBF induction by synaptic activity, is sufficient to reproduce the same form of Volterra kernels that are seen empirically. As such the Balloon model is sufficient to account for the more important nonlinearities observed in evoked fMRI responses. The remainder of this section deals with the validity of the haemodynamic model in terms of the plausibility of the parameter estimates from the previous section. The role of each parameter, in shaping the haemodynamic response, is illustrated in the associated panel in Figure 8 and is discussed in the following subsections.

6.1. The neuronal efficacy (ε)

This represents the increase in perfusion signal elicited by neuronal activity, expressed in terms of event density (i.e. number of evoked transients per second). From a biophysical perspective it is not exceedingly interesting because it reflects both the potency of the stimulus in eliciting a neuronal response and the efficacy of the ensuing synaptic activity to induce the signal. It is interesting to note however that one word per second invokes an increase in normalised rCBF of unity (i.e., in the absence of regulatory effects, a doubling of blood flow over a second). As might be expected changes in this parameter simply modulate the evoked haemodynamic responses (see the first panel in Figure 8).

6.2. Signal decay (τ_s)

This parameter reflects signal decay or elimination. Transduction of neuronal activity into perfusion changes, over a few 100 microns, has a substantial neurogenic component (that may be augmented by electrical conduction up the vascular endothelium). However at spatial scales of several mm it is likely that rapidly diffusing spatial signals mediate increases in rCBF through relaxation of arteriolar smooth muscle. There are a number of candidates for this signal, nitric oxide (NO) being the primary one. It has been shown that the rate of elimination is critical in determining the effective time-constants of haemodynamic transduction (Friston 1995). Our decay parameter had a mean of about 1.54 seconds giving a half-life $t_{1/2} = \tau_s \ln 2 = 1067$ ms. The half-life of NO is between 100 and 1000 ms (Paulson and Newman 1987) whereas that of K^+ is about 5 seconds. Our results are therefore consistent with spatial signalling with NO. It should be remembered that the model signal subsumes all the actual signalling mechanisms employed in the real brain. Increases in this parameter dampen the rCBF response to any input and will also suppress the undershoot (see next subsection) because the feedback mechanisms, that are largely responsible for the undershoot, are selectively suppressed (relative to just reducing neuronal efficacy during signal induction).

6.3. Autoregulation (τ_f)

This parameter is the time-constant of the feedback autoregulatory mechanism whose physiological nature remains unspecified (but see Irikura *et al* 1994). The coupled differential equations Eq(6) and Eq(7) represent a damped oscillator with a resonance frequency of $\varpi = 1/(2\pi\sqrt{\tau_f}) \approx 0.101$ per second. This is exactly the frequency of the vasomotor signal that typically has a period of about 10 seconds. This is a pleasing result that emerges spontaneously from the parameter estimation. The nature of these oscillations can be revealed by increasing the signal decay time constant (i.e. reducing the dampening) and presenting the model with low-level random neuronal input (uncorrelated Gaussian noise with a standard deviation of 1/64) as shown in Figure 9. The characteristic oscillatory dynamics are readily expressed. The effect of increasing the feedback time constant is to decrease the resonance frequency and render the BOLD (and rCBF) response more enduring with a reduction or elimination of the undershoot. The third panel in Figure 8 shows the effect of doubling τ_f.

6.4. Transit time (τ_0)

This is an important parameter that determines the dynamics of the signal. It is effectively resting venous volume divided by resting flow, and in our data is estimated at about one second (0.98 seconds). The transit time through the rat brain is roughly 1.4 seconds at rest and, according to the asymptotic projections for rCBF and volume, falls to 0.73 seconds during stimulation (Mandeville *et al* 1999). In other words it takes about a second for a blood cell to traverse the venous compartment. The effect of increasing mean transit time is to slow down the dynamics of the BOLD signal with respect to the flow changes. The shape of the response remains the same but it is expressed more slowly. In the fourth panel of Figure 8 a doubling of the mean transit time is seen to retard the peak BOLD response by about a second and the undershoot by about 2 seconds.

6.5. Stiffness parameter (α)

Under steady state conditions this would be about 0.38. The mean over voxels considered above was about 0.33. This discrepancy, in relation to steady state levels, is anticipated by the windkessel formulation and is attributable to the fact that volume and flow are in a state of continuous flux during the evoked responses. Recall from Eq(3) that $1/\alpha = \gamma + \beta = 3.03$, in our data. Under the assumption of laminar flow ($\gamma = 2$), $\beta \approx 1$ which is less that the Mandeville *et al* (1999) founds for rats during forepaw stimulation but is certainly in a plausible range. Increasing this parameter increases the degree of nonlinearity in the flow-volume behaviour of the venous balloon that underpins the nonlinear behaviours we are trying to account for. However its direct effect on evoked responses to single stimuli is not very marked. The fifth panel of Figure 8 shows the effects when α is decreased by 50%.

6.6. Resting oxygen extraction (E_0)

This is about 34% and the range observed in our data fit exactly with known values for resting oxygen extraction fraction (between 20% and 55%). Oxygen extraction fraction is a potentially important factor in determining the nature of evoked fMRI responses because it may be sensitive to the nature of the baseline that defines the resting state. Increases in this parameter can have quite profound effects on the shape of the response that bias it towards an early dip. In the example shown (last panel in Figure 8) the resting extraction has been increased to 78%. This is a potentially important observation that may explain why the initial dip has been difficult to observe in all studies. According to the results presented in Figure 8 the initial dip is very sensitive to resting oxygen extraction fraction, which should be high before the dip is expressed. Extraction fraction will be high in regions with very low blood flow, or in tissue with endogenously high extraction. It may be that cytochrome oxidase rich cortex, like the visual cortices, may have a higher fraction and be more like to evidence early dips.

In summary the parameters of the haemodynamic model that best reproduce empirically-derived Volterra kernels are all biologically plausible and lend the model a construct

validity (in relation to the Volterra formulation) and face validity (in relation to other physiological characterisations of the cerebral haemodynamics reviewed in this section). In this extended haemodynamic model nonlinearities, inherent in the Balloon model, have been related directly to nonlinearities in responses. Their role in mediating the post-stimulus undershoot is emphasised less here because the rCBF component can model undershoots.

The conclusions above are based only on data from the auditory cortex and from one subject. There is no guarantee that they will generalise. When submitted in paper form, one of our reviewers thought that it was more important for its conceptual motivation of modelling than for the specific findings. This is a very valid point. We anticipate that the framework presented here will be refined or changed when applied to other data, or the assumptions upon which it is based are confirmed or refuted.

7. Conclusion

In conclusion we have developed an input-state-output model of the haemodynamic response to changes in synaptic activity that combines the Balloon model of flow to BOLD signal coupling and a dynamical model of the transduction of neuronal activity into perfusion changes. This model has been characterised in terms of its Volterra kernels and easily reproduces empirical kernels with parameters that are biologically plausible. This means that the nonlinearities inherent in the Balloon model are sufficient to account for haemodynamic refractoriness and other nonlinear aspects of evoked responses in fMRI.

8. Appendix

Volterra kernels represent a generic and important characterisation of the invariant aspects of a nonlinear system (see Bendat 1990). This appendix describes the nature of these kernels and how they are obtained given the differential equations describing the evolution of the state variables. Consider the single input-single output (SISO) system

$$\begin{aligned} \dot{X}(t) &= f(X, u(t)) \\ y(t) &= \lambda(X(t)) \end{aligned} \qquad \text{A.1}$$

where, for the haemodynamic model, $X = \{x_1, x_2, x_3, x_4\}^T = \{s, f_{in}, v, q\}^T$ with

$$\begin{aligned} \dot{x}_1 &= f_1(X, u(t)) = \varepsilon u(t) - \frac{x_1}{\tau_s} - \frac{x_2 - 1}{\tau_f} \\ \dot{x}_2 &= f_2(X, u(t)) = x_1 \\ \dot{x}_3 &= f_3(X, u(t)) = \frac{1}{\tau_0}\left(x_2 - f_{out}(x_3, \alpha)\right) \\ \dot{x}_4 &= f_4(X, u(t)) = \frac{1}{\tau_0}\left(x_2 \frac{E(x_2, E_0)}{E_0} - f_{out}(x_3, \alpha)\frac{x_4}{x_3}\right) \end{aligned}$$

and $$y(t) = \lambda(X(t)) = V_0\left(k_1(1 - x_4) + k_2(1 - x_4 / x_3) + k_3(1 - x_3)\right)$$

The Volterra series expresses the output $y(t)$ as a nonlinear convolution of the neuronal inputs $u(t)$, critically without reference to the state variables $X(t)$. This series can be considered a nonlinear convolution that obtains from a functional Taylor expansion of $y(t)$ about $X(0) = X_0 = [0, 1, 1, 1]^T$ and $u(t) = 0$

$$\begin{aligned} y(t) &= \kappa_0(t) + \sum_{i=1}^{\infty} \int_0^t \ldots \int_0^t \kappa_i(t, \sigma_1, \ldots \sigma_i) u(\sigma_1) \ldots u(\sigma_i) d\sigma_1 \ldots d\sigma_i \\ \kappa_i(t, \sigma_1, \ldots \sigma_i) &= \frac{\partial^i y(t)}{\partial u(\sigma_1) \ldots u(\sigma_i)} \end{aligned} \qquad \text{A.2}$$

where κ_i is the ith, generally time-dependent, kernel. The Taylor expansion of $\dot{X}(t)$ about X_0 and $u(t) = 0$

$$\dot{X}(t) \approx f(X_0,0) + \frac{\partial f(X_0,0)}{\partial X}(X - X_0) + \frac{\partial^2 f(X_0,0)}{\partial X \partial u}(X - X_0)u + \frac{\partial f(X_0,0)}{\partial u}u$$

has a bilinear form following a change of variables (equivalent to adding an extra state variable $x_0(t) = 1$)

$$\begin{aligned}
&\dot{X}'(t) \approx AX' + BX'u \\
&X' = \begin{bmatrix} 1 \\ X \end{bmatrix} \\
&A = \begin{bmatrix} 0 & 0 \\ \left(f(X_0,0) - \dfrac{\partial f(X_0,0)}{\partial X} X_0 \right) & \dfrac{\partial f(X_0,0)}{\partial X} \end{bmatrix} \\
&B = \begin{bmatrix} 0 & 0 \\ \left(\dfrac{\partial f(X_0,0)}{\partial u} - \dfrac{\partial^2 f(X_0,0)}{\partial X \partial u} X_0 \right) & \dfrac{\partial^2 f(X_0,0)}{\partial X \partial u} \end{bmatrix}
\end{aligned} \qquad \text{A.3}$$

This formulation is important because the Volterra kernels of bilinear systems have closed-form expressions. The existence of these closed-form expressions is due to the fact that the iterated integrals associated with the system's Generating Series can be expressed in terms of the generalised convolution integrals, of which the Volterra series is comprised (Fliess *et al* 1983). Here we take a more heuristic approach and consider the solution to A.2 and its derivatives with respect to the inputs $u(t)$

$$X'(\Delta t) \approx e^{\Delta t(A+Bu(0))}X'(0) \quad \Rightarrow \quad X'(T\Delta t) \approx \prod_{j=T-1}^{0} e^{\Delta t(A+Bu(j\Delta t))}X'(0), \qquad \Delta t \to 0$$

$$\frac{\partial^i X'(T\Delta t)}{\partial u(\tau_1 \Delta t)\ldots u(\tau_i \Delta t)} = \prod_{j=T-1}^{\tau_i+1} e^{\Delta t(A+Bu(j\Delta t))} B \prod_{j=\tau_i}^{\tau_{i-1}+1} e^{\Delta t(A+Bu(j\Delta t))} \ldots B \prod_{j=\tau_1}^{0} e^{\Delta t(A+Bu(j\Delta t))} X'(0)$$

The kernels associated with the state variables X' are these derivatives evaluated at $u(t)=0$

$$\chi_i(t,\sigma_1,\ldots\sigma_i) = \frac{\partial^i X'(t)}{\partial u(\sigma_1)\ldots u(\sigma_i)} = e^{(t-\sigma_i)A} B e^{(\sigma_i-\sigma_{i-1})A} \ldots B e^{\sigma_1 A} X'(0)$$

i.e.
$$\chi_0(t) = e^{tA} X'(0)$$
$$\chi_1(t,\sigma_1) = e^{(t-\sigma_1)A} B e^{\sigma_1 A} X'(0)$$
$$\chi_2(t,\sigma_1,\sigma_2) = e^{(t-\sigma_2)A} B e^{(\sigma_2-\sigma_1)A} B e^{\sigma_1 A} X'(0)$$
$$\chi_2(t,\sigma_1,\sigma_2,\sigma_3) = \ldots$$

The kernels associated with the output $y(t)$ follow from the chain rule

$$\kappa_0(t) = \lambda\big(\chi_0(t)\big)$$
$$\kappa_1(t,\sigma_1) = \frac{\partial\lambda\big(\chi_0(t)\big)}{\partial X'}\chi_1(t,\sigma_1)$$
$$\kappa_2(t,\sigma_1,\sigma_2) = \frac{\partial\lambda\big(\chi_0(t)\big)}{\partial X'}\chi_2(t,\sigma_1,\sigma_2) + \chi_1(t,\sigma_1)^{\mathrm{T}}\frac{\partial^2\lambda\big(\chi_0(t)\big)}{\partial X'^2}\chi_1(t,\sigma_2)$$
$$\kappa_2(t,\sigma_1,\sigma_2,\sigma_3) = \ldots$$

If the system is fully nonlinear, as in this case then the kernels can be considered local approximations. In other words the kernels are valid for inputs (i.e. neuronal activations) of a reasonable magnitude.

9. Acknowledgements

This work was funded by the Wellcome trust and the MRC. We would like to thank Gary Green for guidance and support in understanding and using Volterra series.

10. References

JS Bendat. (1990) Nonlinear System Analysis and Identification from Random Data. John Wiley and Sons, New York USA

GM Boynton, SA Engel, GH Glover, and DJ Heeger. (1996) Linear systems analysis of functional magnetic resonance imaging in human V1. *J. Neurosci.* **16**,4207-4221

RB Buxton and LR Frank. (1997) A model for the coupling between cerebral blood flow and oxygen metabolism during neural stimulation. *J. Cereb. Blood Flow Metab*. **17**, 64-72

RB Buxton, EC Wong, and LR Frank. Dynamics of blood flow and oxygenation changes during brain activation: The Balloon model. (1998) *MRM* **39**, 855-864

M Fliess, M Lamnabhi and F Lamnabhi-Lagarrigue (1983) An algebraic approach to nonlinear functional expansions. *IEEE Trans. Circuits Syst.* **30**, 554-570

PT Fox and ME Raichle (1986) Focal physiological uncoupling of cerebral blood flow and oxidative metabolism during somatosensory stimulation in human subjects. *Proc Natl Acad Sci U S A*. Feb;**83(4)**:1140-4.

KJ Friston, P Jezzard, and R Turner (1994) Analysis of functional MRI time series. *Human Brain Map*. **1**,153-171

KJ Friston (1995) Regulation of rCBF by diffusible signals: An analysis of constraints on diffusion and elimination *Hum Brain Mapp.* **3**, 56-65

KJ Friston, O Josephs, G Rees, and R Turner. (1998) Nonlinear event-related responses in fMRI. *MRM* **39**, 41-52

K Grill-Spector, R Malach (2001) fMR-adaptation: a tool for studying the functional properties of human cortical neurons. *Acta Psychol (Amst).* **107(1-3)**:293-321.

RL Grubb, ME Rachael, JO Euchring, and MM Ter-Pogossian. (1974) The effects of changes in PCO_2 on cerebral blood volume, blood flow and vascular mean transit time. *Stroke* **5**, 630-639

DJ Heeger and D Ress (2002) What does fMRI tell us about neuronal activity? *Nat Rev Neurosci.* **3(2)**:142-51.

RD Hoge, J Atkinson, B Gill, GR Crelier, S Marrett and GB Pike (1999) Linear coupling between cerebral blood flow and oxygen consumption in activated human cortex. *Proc. Natl. Acad. Sci. 96*, 9403-9408

K Irikura, KI Maynard, and MA Moskowitz (1994) Importance of nitric oxide synthase inhibition to the attenuated vascular responses induced by topical l-nitro-arginine during vibrissal stimulation. *J. Cereb. Blood Flow Metab*. **14**, 45-48

C Kennedy, MH Des Rosiers, O Sakurada, M Shinohara, M Reivich, JW Jehle, L Sokoloff (1976) Metabolic mapping of the primary visual system of the monkey by means of the autoradiographic [14C]deoxyglucose technique. *Proc Natl Acad Sci U S A.* **73(11)**:4230-4.

U Lindauer, G Royl, C Leithner, M Kuhl, L Gold, J Gethmann, M Kohl-Bareis, A Villringer, U Dirnagl (2001) No evidence for early decrease in blood oxygenation in rat whisker cortex in response to functional activation. *Neuroimage*. Jun;**13(6 Pt 1)**:988-1001.

NK Logothetis, J Pauls, M Augath, T Trinath, A Oeltermann (2001) Neurophysiological investigation of the basis of the fMRI signal. *Nature* **412(6843)**:150-7.

JB Mandeville, JJ Marota, C Ayata, G Zararchuk, MA Moskowitz, B Rosen and RM Weisskoff (1999) Evidence of a cerebrovascular postarteriole windkessel with delayed compliance. *J. Cereb. Blood Flow Metab*. **19**, 679-689

J Mayhew, D Hu, Y Zheng, S Askew, Y Hou, J Berwick, PJ Coffey, and N Brown (1998) An evaluation of linear models analysis techniques for processing images of microcirculation activity *NeuroImage* **7**, 49-71

KJ Friston, A Mechelli, R Turner, CJ Price (2000) Nonlinear responses in fMRI: the Balloon model, Volterra kernels, and other hemodynamics. *Neuroimage*. **12(4)**:466-77.KL Miller, WM Luh, TT Liu, A Martinez, T Obata, EC Wong, LR Frank and RB Buxton (2000) Characterizing the dynamic perfusion response to stimuli of short duration. *Proc. ISRM* **8**, 580

OB Paulson, and EA Newman (1987) Does the release of potassium from astrocyte endfeet regulate cerebral blood? *Science* **237**, 896-898

CS Roy, CS Sherrington (1890) On the regulation of the blood supply of the brain. *J Physiol Lond* **11**:85-108.

J Talairach, and P Tournoux. (1988) *A Co-planar stereotaxic atlas of a human brain.* Thieme, Stuttgart.

I Vanzetta and A Grinvald (1999) Increased cortical oxidative metabolism due to sensory stimulation: implications for functional brain imaging. *Science*. **286(5444)**:1555-8.

AL Vazquez, and DC Noll DC. (1996) Non-linear temporal aspects of the BOLD response in fMRI. *Proc. Int. Soc. Mag. Res. Med.* (Vol **3**), S1765

KJ Worsley, and KJ Friston.(1995) Analysis of fMRI time-series revisited - again *NeuroImage* **2**,173-181

Y Zheng, J Martindale, D Johnston, M Jones, J Berwick, J Mayhew (2002) A model of the hemodynamic response and oxygen delivery to brain. *Neuroimage* **16(3 Pt 1)**:617-37.

11. Figure Legends

Figure 1
Top panel: Simulated responses to a pair of words (bars) one second apart, presented together (solid line) and separately (broken lines) based on the kernels shown in Figure 4. Lower panel: The response to the second word when presented alone (broken line as above) and when preceded by the first (solid line). The latter obtains by subtracting the response to the first word from the response to both. The difference reflects the effect of the first word on the response to the second.

Figure 2
Schematic illustrating the organisation of the haemodynamic model. This is a fully nonlinear single input $u(t)$, single output $y(t)$ state model with four state variables s, f, v and q. The form and motivation for the changes in each state variable, as functions of the others, is described in the main text.

Figure 3
Dynamics of the haemodynamic model. Upper left panel: The time-dependent changes in the neuronally induced perfusion signal that causes an increase in blood flow. Lower left panel: The resulting changes in normalised blood flow (*f*). Upper right panel: The concomitant changes in normalised venous volume (*v*) (solid line) and normalised deoxyhaemoglobin content (*q*) (broken line). Lower right panel: The percent change in BOLD signal that is contingent on *v* and *q*. The broken line is inflow normalised to the same maximum as the BOLD signal. This highlights the fact that BOLD signal lags the rCBF signal by about a second.

Figure 4
Voxels used to estimate the parameters of the haemodynamic model shown in Figure 2. This is a SPM{F} testing for the significance of the first and second order kernel coefficients in the empirical analysis and represents a maximum intensity projection of a statistical process of the F ratio, following a multiple regression analysis at each voxel. This regression analysis estimated the kernel coefficients after expanding them in terms of a small number of temporal basis functions (see Friston *et al* 1998 for details). The format is standard and provides three orthogonal projections in the standard space conforming to that described in Talairach and Tournoux (1988). The grey scale is arbitrary and the SPM{F} has been thresholded to show the 128 most significant voxels.

Figure 5
The first and second order Volterra kernels based on parameter estimates from a voxel in the left superior temporal gyrus at -56, -28, 12mm. These kernels can be thought of as a second order haemodynamic response function. The first order kernels (upper panels) represent the (first order) component usually presented in linear analyses. The second order kernels (lower panels) are presented in image format. The colour scale is arbitrary; white is positive and black is negative. The left-hand panels are kernels based on parameter estimates from the analysis described in Figure 4. The right hand panels are the kernels associated with the haemodynamic model using parameter estimates that best match the empirical kernels.

Figure 6
Functions implied by the [mean] haemodynamic model parameters over the voxels shown in Figure 4. Upper left panel: Outflow as a function of venous volume $f_{out}(v,\alpha)$. Upper right panel: oxygen extraction as a function of inflow. The solid line is extraction *per se* $E(f_{in}, E_0)$ and the broken line is the net normalised delivery of deoxyhaemoglobin to the venous compartment $f_{in}E(f_{in}, E_0)/E_0$. Lower panel: This is a plot of the nonlinear function of volume and deoxyhaemoglobin that represents BOLD signal $y(t) = \lambda(v, q, E_0)$.

Figure 7

Histograms of the distribution of the 6 free parameters of the haemodynamic model estimated over the voxels shown in Figure 3. The number in brackets at the top of each histogram is the mean value for the parameters in question: neuronal efficacy is ε, signal decay is τ_s, autoregulation is τ_f, transit time is τ_0, stiffness is α and oxygen extraction is E_0.

Figure 8
The effects of changing the model parameters on the evoked BOLD response. The number in brackets at the top of each graph is the factor applied to the parameter in question. Solid lines correspond to the response after changing the parameter and the broken line is the response for the original parameter values (the mean values given in Figure 7): neuronal efficacy is ε, signal decay is τ_s, autoregulation is τ_f, transit time is τ_0, stiffness is α and oxygen extraction is E_0.

Figure 9
Simulated response to a noisy neuronal input (standard deviation 1/64 and mean of 0) for a model with decreased signal decay (i.e. less dampening). The model parameters were the same as in the Figure 3 with the exception of τ_s which was increased by a factor of 4. The characteristic 0.1 Hz oscillations are very similar to the oscillatory vasomotor signal seen in optical imaging experiments.

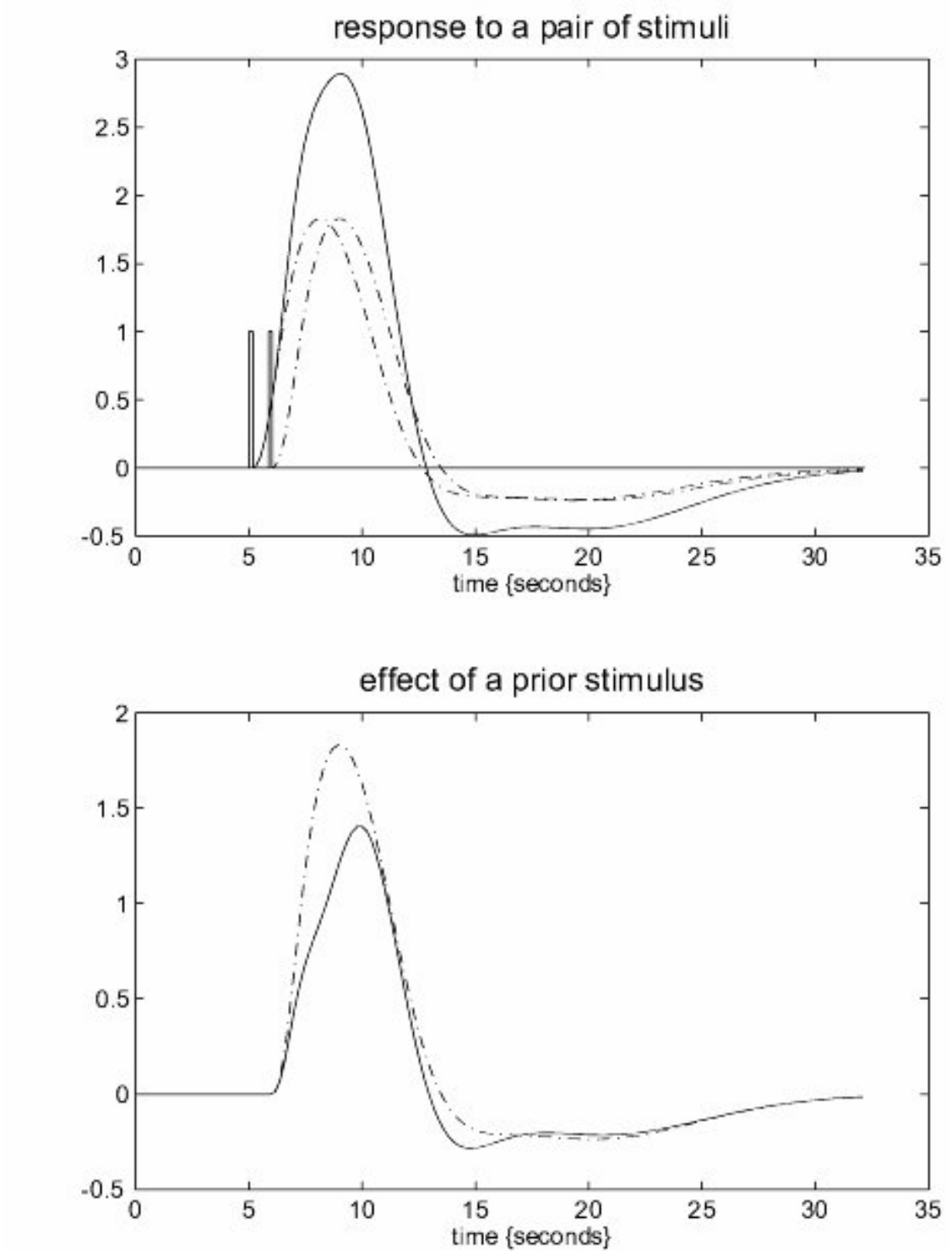
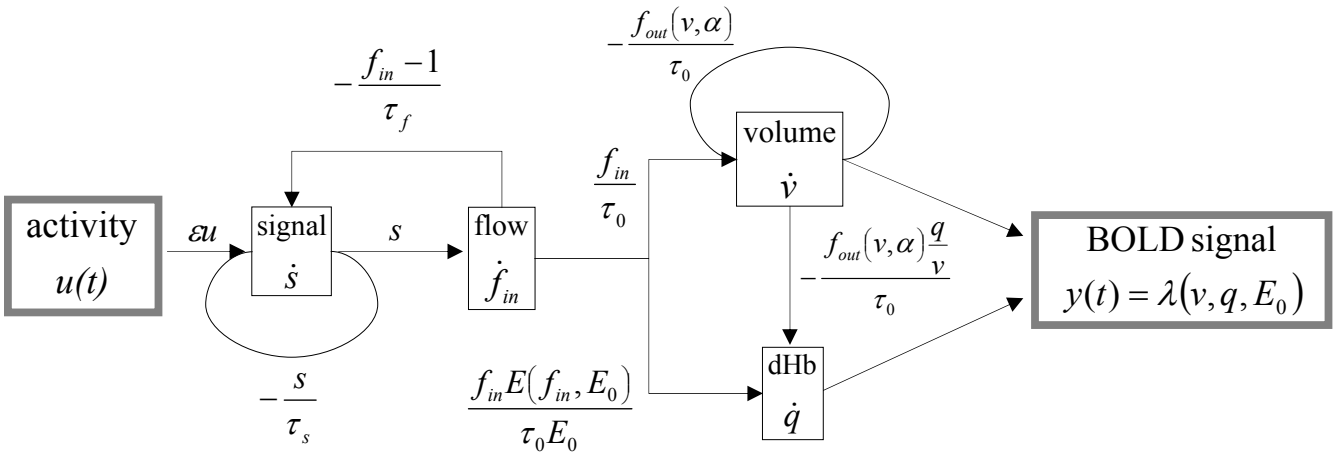
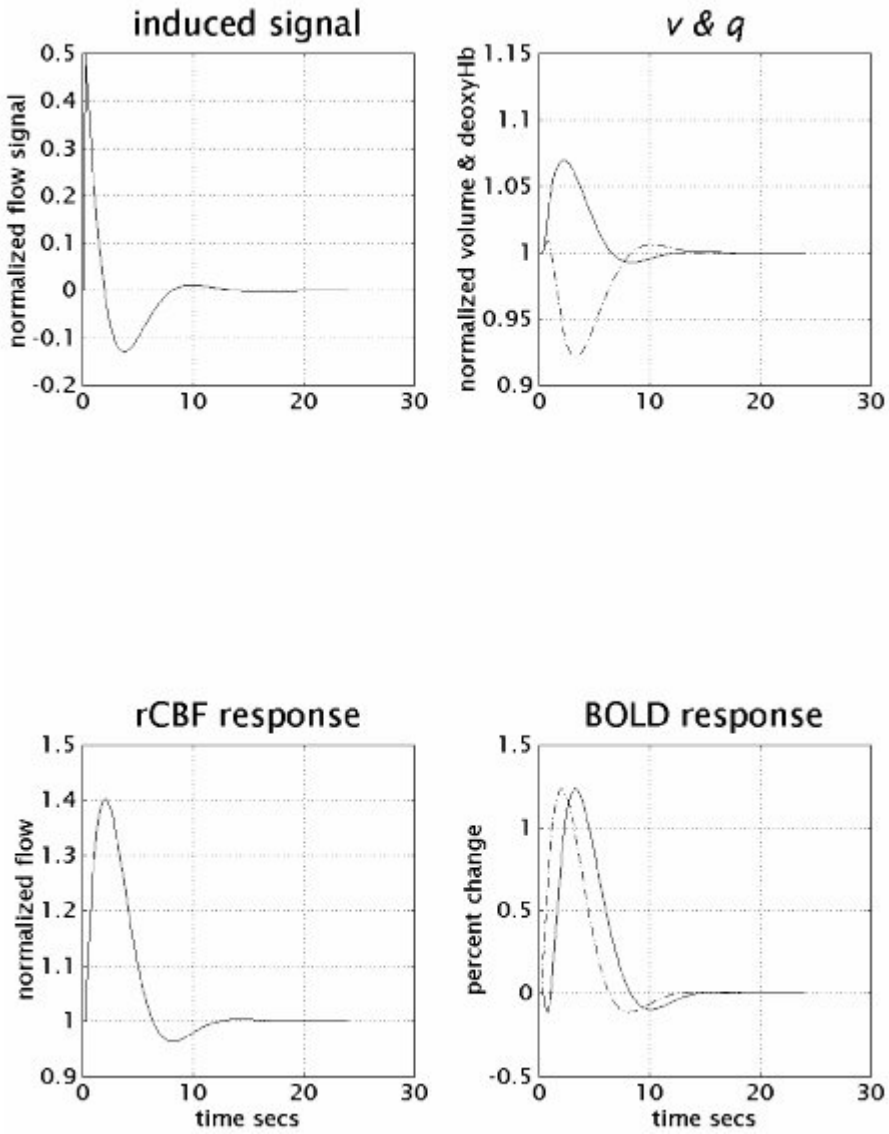

response to a pair of stimuli
3
2.5
2
1.5
1
0.5
0
-0.5
0
5
10
15
20
25
30
35
time {seconds}
effect of a prior stimulus
2
1.5
1
0.5
0
-0.5
0
5
10
15
20
25
30
35
time {seconds}

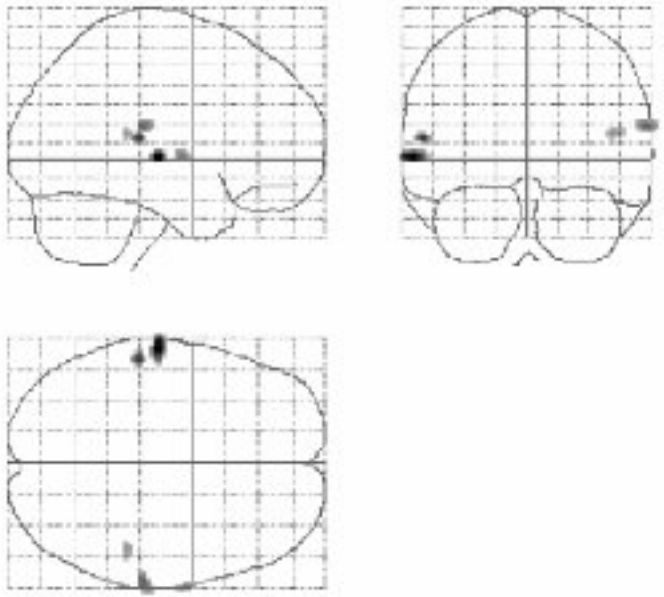
Figure 1

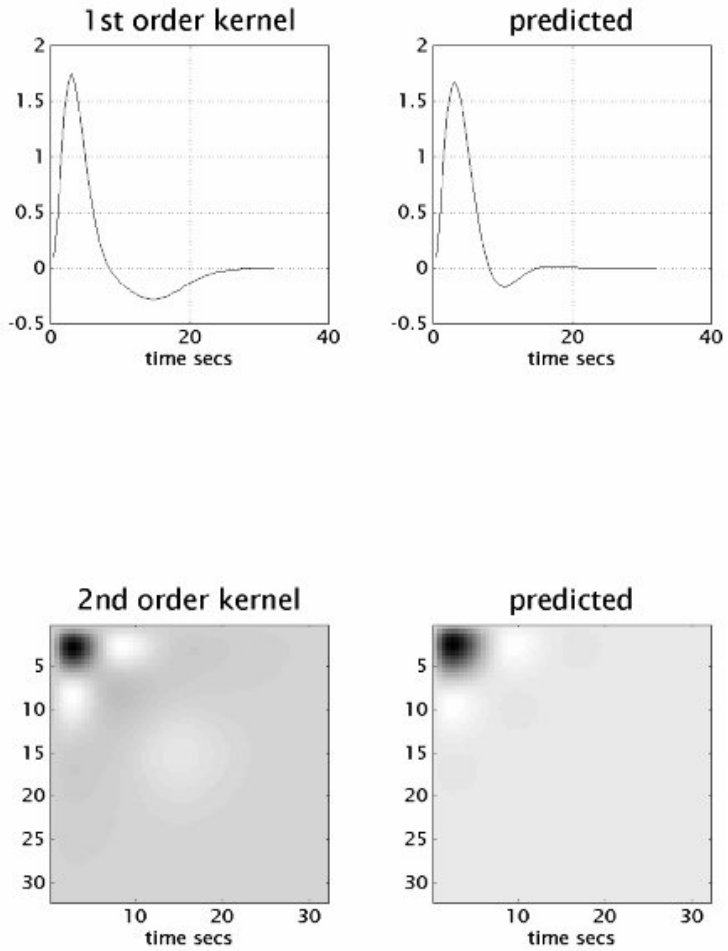
Figure 2

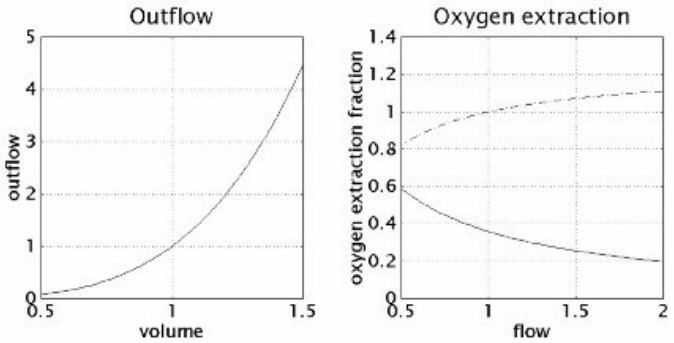

activity
u(t)
εu
signal
ṡ
s
flow
ḟ_in
volume
v̇
dHb
q̇
BOLD signal
y(t) = λ(v, q, E_0)
−(f_in − 1)/τ_f
−s/τ_s
f_in/τ_0
−f_out(v,α)/τ_0
−f_out(v,α) q/v / τ_0
f_in E(f_in, E_0)/(τ_0 E_0)

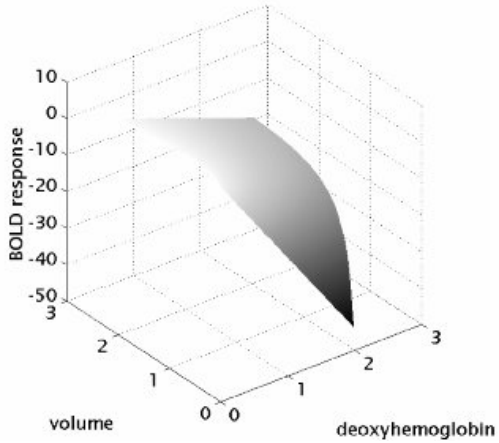

induced signal
normalized flow signal
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
0
10
20
30
v & q
normalized volume & deoxyHb
1.15
1.1
1.05
1
0.95
0.9
0
10
20
30
rCBF response
normalized flow
1.5
1.4
1.3
1.2
1.1
1
0.9
0
10
20
30
time secs
BOLD response
percent change
1.5
1
0.5
0
-0.5
0
10
20
30
time secs

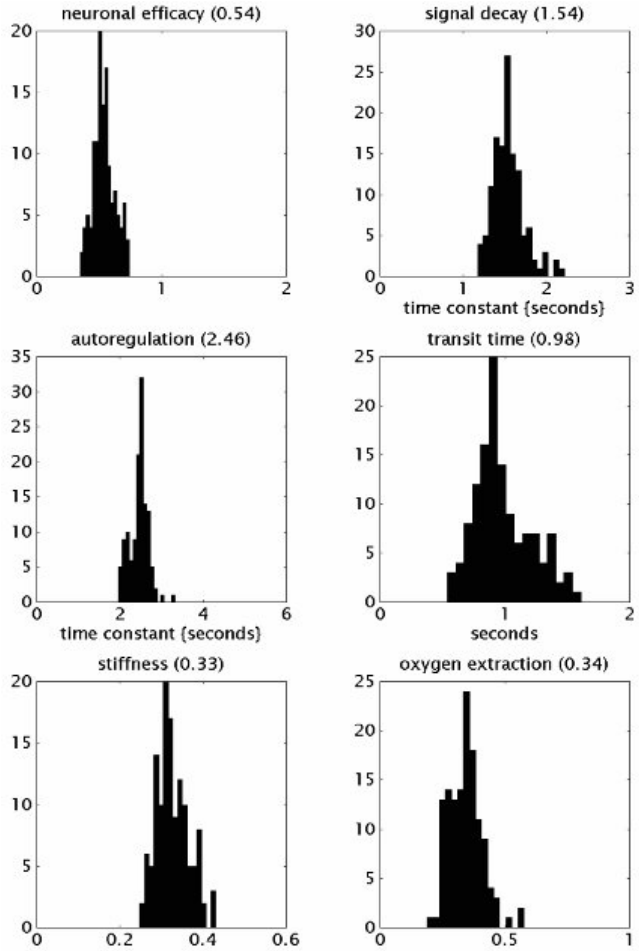
Voxels analyzed

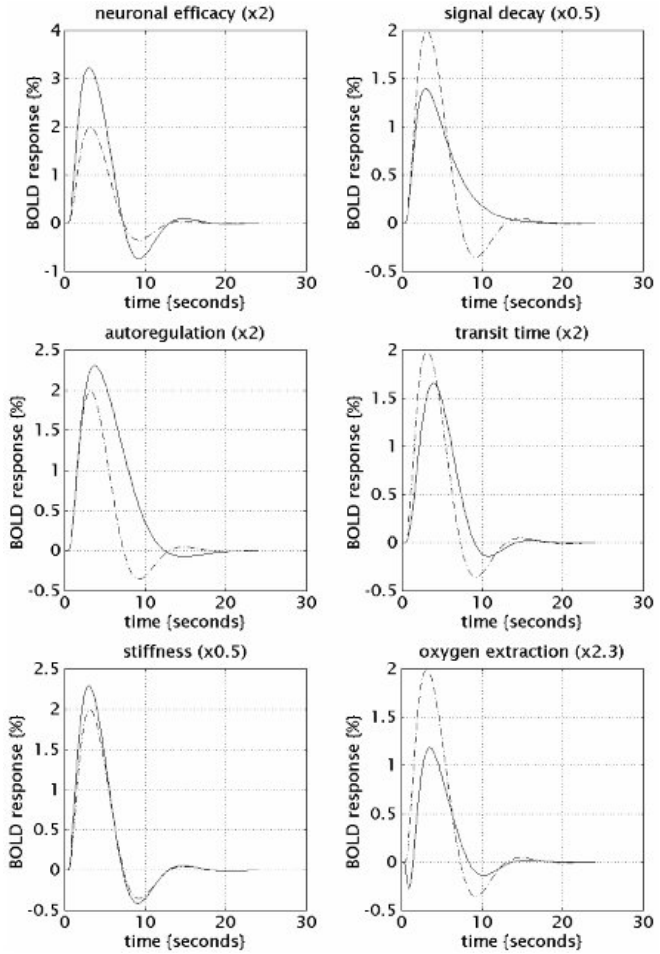


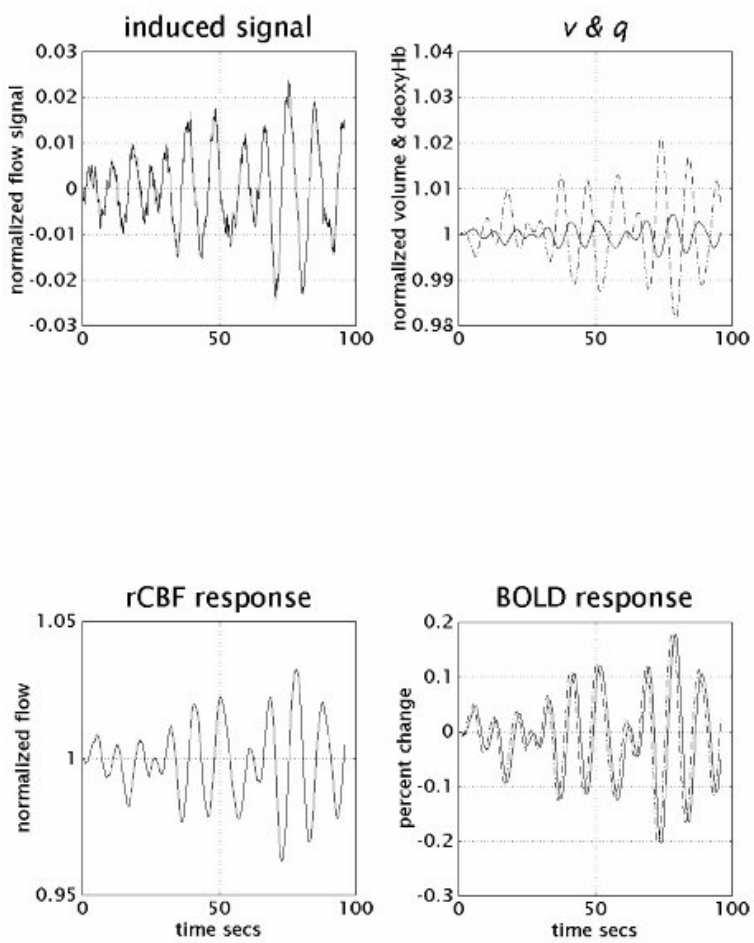

1st order kernel
predicted
2
1.5
1
0.5
0
-0.5
0
20
40
time secs
2nd order kernel
predicted
5
10
15
20
25
30
10
20
30
time secs


Outflow
outflow
volume
0
1
2
3
4
5
0.5
1
1.5
Oxygen extraction
oxygen extraction fraction
flow
0
0.2
0.4
0.6
0.8
1
1.2
1.4
0.5
1
1.5
2


BOLD response
10
0
-10
-20
-30
-40
-50
volume
deoxyhemoglobin
0
1
2
3
0
1
2
3


neuronal efficacy (0.54)
signal decay (1.54)
time constant {seconds}
autoregulation (2.46)
time constant {seconds}
transit time (0.98)
seconds
stiffness (0.33)
oxygen extraction (0.34)


neuronal efficacy (x2)
signal decay (x0.5)
autoregulation (x2)
transit time (x2)
stiffness (x0.5)
oxygen extraction (x2.3)
BOLD response [%]
time {seconds}


induced signal
normalized flow signal
0.03
0.02
0.01
0
-0.01
-0.02
-0.03
0
50
100
v & q
normalized volume & deoxyHb
1.04
1.03
1.02
1.01
1
0.99
0.98
0
50
100
rCBF response
normalized flow
1.05
1
0.95
0
50
100
time secs
BOLD response
percent change
0.2
0.1
0
-0.1
-0.2
-0.3
0
50
100
time secs