Referee 2.

We thank the referee for the careful reading of the manuscript.

Regarding the more general practical utility of the proposed sampling protocol, while experimental relaxation data collected at four magnetic field strengths yields 6859 suitably filtered combinations of bootstrap samples, as noted by the first reviewer, the robustness of the statistical analysis may appreciably decline when this value drops to 343 for three magnetic field strengths, and presumably will decrease significantly further when it drops to only 27 for data from two magnetic field strengths.

> **Response**: The reduction to three fields has been discussed in the response to referee 1. Further reduction to two fields would not allow any reasonable resampling.

A key step in the proposed joint refinement process calculates each of the averaged dynamical parameters by summing over the estimates obtained from each of the five spectral density representations being utilized, as weighted by how often each of these five models have been selected (Eq. 10). A potential concern over this approach arises from the fact that while the same set of symbols (τ_m, S_f^2,τ_f,S_s^2,τ_s) are utilized in each of the five dynamical models used, the functional significance of each symbol is defined within the context of the specific equation being used.

> **Response**: The referee is correct in the sense that if only one internal time scale is fit, for example, τ_s, then the optimized value of the parameter will average over both fast and slow time scales (in some complex manner), whereas if both τ_f and τ_s are included in the model, then some partitioning of the time scales occurs. This issue is exactly what the bootstrap aggregation procedure addresses. That said, care must be taken in the interpretation of the fitted parameters for different models when aggregating results. In the present application, models that incorporated a single internal correlation time were partitioned between τ_f (model 2) and τ_s (model 3) based on an empirical criterion, as described in the paper. One can imagine situations in which deciding how to perform the averaging between model parameters would be a more difficult question.

Each of these five model equations that are used to represent the spectral density function is capable of accurately fitting only a small subset of the physically plausible spectral density curves. Systematic bias can potentially arise not only with respect to a given dynamics parameter being utilized in distinct model representations but also as a result of the inadequacy with which each of the five model spectral density equations are capable of representing the physical dynamics of the system. While such biasing effects are surely diminished for Model 4 and 5 which incorporate four and five adjustable parameters, respectively, more promising might be the utilization of alternative model equations for the spectral density function that can more robustly represent the range of motion occurring in protein molecules which utilize a smaller set of adjustable parameters for optimization against experimental relaxation data.

> **Response.** The referee raises important question concerning the nature of the models used to fit relaxation (or any) data. The merits and limitations of the "model-free" approach to

analyzing spin relaxation data have been discussed beginning with Lipari and Szabo in their original papers and many others subsequently. A number of alternative models for the spectral density function have been proposed, including a number of interesting distribution functions for correlation times, and molecular dynamics simulations are becoming more capable of directly estimating relaxation rate constants. The introductory paragraph of the paper introduced some of these alternative approaches, but was not comprehensive. This paragraph has been expanded in the revision to include some additional recent work in this arena. Nonetheless, the model-free strategy remains the most widespread approach used for analysis of spin relaxation data. The present paper addresses a weakness in this approach: model selection error, and hence will be useful to a large number of researchers. At the same time, bootstrap aggregation is a general approach for treating model selection error and can be applied in the context of other approaches for analyzing relaxation data.